Dempster-Shafer Theory

Jonathan Weisberg
University of Toronto

Northern Institute of Philosophy
June 16, 2010

Outline

1. Motivations
2. Statics
3. Dynamics
4. Taxonomy
5. Decisions

Motivations
Statics
Dynamics
Taxonomy
Decisions
References

Motivations

Motivations

1. Representing ignorance
2. The problem of priors
3. Symmetric treatment of prior belief & evidence
4. Representing evidence:
   - Evidential basis
   - Weights of evidence
   - Uncertain evidence

Two Simple Examples

A vase appears to be a Ming. Let $\Omega = \{w_1, w_2\}$ represent the space of possibilities that it is genuine (w_1) or fake (w_2).

A State of Ignorance

Let $bel : 2^\Omega \to [0, 1]$ be the function given by

	$\varnothing$	$\{w_1\}$	$\{w_2\}$	Ω
bel	0	0	0	1

An expert then attests that it is probably fake:

A Simple Support Function

	$\varnothing$	$\{w_1\}$	$\{w_2\}$	Ω
bel_E	0	0	s	1

$s > 0$

Representing Ignorance

Shafer (1976: p. 22) advertises belief functions as representing ignorance better than probability functions.

- The problem with probability functions is well known.
  - Ignorance is represented by uniform distributions.
  - But refining the space of possibilities yields inconsistency.
- Belief functions avoid this problem.
  - Ignorance is represented by *vacuous* belief functions.
  - Refining the space preserves previous assignments.

Refined Ignorance

Suppose we think to distinguish early Ming (w_1^1) from late Ming (w_1^2). Then $\Omega = \{w_1^1, w_1^2, w_2\}$ and ignorance is represented:

	$\varnothing$	$\{w_i^{(j)}\}$	$\{w_i^{(j)}, w_k^{(l)}\}$	Ω
bel	0	0	0	1

The Problem of Priors

As a sort of corollary, the problem of priors is easily answered in Dempster-Shafer Theory.

- Your initial degrees of belief should be vacuous: 0 everywhere but the tautology.
- At any later time, your degrees of belief should be the result of combining the vacuous belief function with your total evidence.

Symmetry Between Prior Belief & Evidence

Shafer (1976: p. 25) advertises DST as treating evidence and prior opinion symmetrically.

- Belief-states and evidence are represented by the same sorts of mathematical objects, belief functions.
- Updating is done by combining your priors (*bel*) and your new evidence (bel_E) via a commutative operation, $\oplus$.

$$bel' = bel \oplus bel_E = bel_E \oplus bel$$

- Corollaries:
  - Old and new evidence are treated the same.
    - bel_E is incorporated the same way as the old evidence that generated *bel*.
  - Updating is *commutative*, or *order-invariant*.
    - Compare the classic complaint about Jeffrey's Rule.

Representing Evidence

Shafer's theory is about degrees of belief *based on evidence*. As such, it has at least three selling points:

1. Representing one's evidential basis: in a sense, degrees of belief are nothing more than the sum of one's evidence.
   - As we'll see, combining the vacuous belief function (bel_0) with any other is always neutral: $bel_0 \oplus bel = bel$. So

$$\begin{aligned} bel' &= bel_0 \oplus bel_{E_1} \oplus \ldots \oplus bel_{E_n} \\ &= bel_{E_1} \oplus \ldots \oplus bel_{E_n} \end{aligned}$$

   - In fact, one can often decompose a belief function into the evidence upon which it is based.
   - There are limitations, of course; more on that when we discuss Shafer's Theorem 5.2.

Representing Evidence

2. Weights of evidence: the theory lends itself to a natural way of measuring the weight of evidence for A.
   - Suppose we have two pieces of evidence for A, bel_{E_1} and bel_{E_2}.
   - Assume that the weights of evidence underlying these two pieces of evidence combines additively: $w_3 = w_1 + w_2$.
   - Then we can derive (with some "innocuous" assumptions):

$$bel(A) = 1 - e^{-w(A)}$$

   - This relation has some intuitively nice features, and supports some interesting theorems/conjectures in DST.

Representing Evidence

3 DST allows us to represent uncertain evidence, and simply.

- The belief function

$$bel(A) = \begin{cases} 0 & \text{if } E \nsubseteq A \\ s & \text{if } E \subseteq A \text{ but } A \neq \Omega \\ 1 & \text{if } A = \Omega \end{cases}$$

  represents evidence that supports degree of belief s in E.
- Compare the classic complaints about conditioning:
  - Evidence must be certain.
  - Evidence must have a pre-existing degree of belief.
- What about Jeffrey's rule? It "…still treats the old and new evidence asymmetrically".
  - Is this a complaint about commutativity?
  - If so, I'd say (Lange, 2000; Wagner, 2002) resolve that worry.

The Horse's Mouth

"For what reasons are degrees of belief required to satisfy the conditions imposed? [. . .] I do not deny the possibility of a theory superior to the theory of belief functions. I believe, though, that the superiority of one theory of probability judgment to another can be demonstrated only by a preponderance of examples where the best analysis using the other." (Shafer 1981a: 15)

Statics

Belief Functions

Definition: *Belief Function*

A function $bel : 2^{\Omega} \to [0, 1]$ is a belief function iff

(B1) $bel(\varnothing) = 0$

(B2) $bel(\Omega) = 1$

(B3) For all $A_1, \ldots, A_n \subseteq \Omega$,

$$bel(A_1 \cup \ldots \cup A_n) \geq \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} bel\left(\bigcap_{i \in I} A_i\right)$$

- (B1) and (B2) are the same as in probability theory.
- So what's the deal with (B3)?

Inclusion-Exclusion

Recall the inclusion-exclusion rule from probability theory:

$$
\begin{aligned}
p(A \cup B) &= p(A) + p(B) \\
&\quad - p(A \cap B) \\
p(A \cup B \cup C) &= p(A) + p(B) + p(C) \\
&\quad - p(A \cap B) - p(A \cap C) - p(B \cap C) \\
&\quad + p(A \cap B \cap C) \\
&\vdots \\
p(A_1 \cup \ldots \cup A_n) &= \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} p\left(\bigcap_{i \in I} A_i\right)
\end{aligned}
$$

$=$ vs. $\geq$

So (B3) is DST's analogue of the inclusion-exclusion rule:

$$p(A_1 \cup \ldots \cup A_n) = \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} p\left(\bigcap_{i \in I} A_i\right)$$

vs.

$$bel(A_1 \cup \ldots \cup A_n) \geq \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} bel\left(\bigcap_{i \in I} A_i\right)$$

- Recall that the inclusion-exclusion principle can replace the additivity axiom of probability theory.
- So the difference between DST and probability theory comes down to replacing a single $=$ with a $\geq$!

But Oh, What a Difference…

A striking (and annoying) novelty of DST is that the values of the atoms do not determine the whole distribution.

- If $\Omega = \{w_1, w_2, w_3\}$, the following is a belief function:

	$\emptyset$	$\{w_i\}$	$\{w_i, w_j\}$	Ω
bel	0	1/4	3/4	1

$i \neq j$

- So is the vacuous function:

	A	Ω
bel	0	1

$A \neq \Omega$

- Another handy trick you'll miss:

$$p(\overline{A}) = 1 - p(A)$$

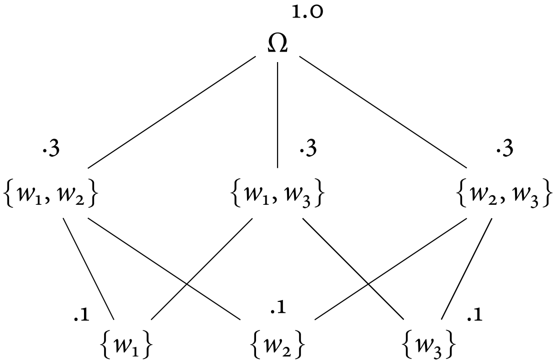
In general, we say that belief functions are *superadditive*:

$$bel(A \cup B) \geq bel(A) + bel(B), \;\; A \cap B = \emptyset$$

How Annoying!

As a result, visual representation is messier:

- We can't visualize belief distributions as "muddy" venn diagrams, in the manner of (van Fraassen, 1989).
- We can use lattices instead:



Mass Functions

Definition: *Mass Function*

A function $m : 2^{\Omega} \to [0, 1]$ is a mass function iff

(M1) $m(\varnothing) = 0$

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

Representation Theorem

Given a mass function m,

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

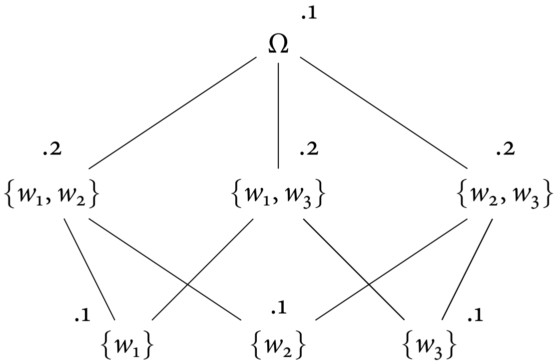
is a belief function.

If Ω is finite and *bel* is a belief function, there is a unique mass function m,

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

Understanding Mass Functions

- Pictorially, mass functions are like probability distributions over the lattice:



- Intuitively, mass is the amount of “belief that one commits exactly to A, not the total belief that one commits to A.”

Commonality Functions

Definition: Commonality Function

If m is a mass function, then its commonality function is

$$Q(A) = \sum_{A \subseteq B,\, B \subseteq \Omega} m(B)$$

Representation Theorem

Given *bel* and its corresponding Q,

$$\begin{aligned} bel(A) &= \sum_{B \subseteq \overline{A}} (-1)^{|B|} Q(B) \\ Q(A) &= \sum_{B \subseteq A} (-1)^{|B|} bel(\overline{B}) \end{aligned}$$

Plausibility Functions

Definition: Plausibility Function

If *bel* is a belief function, its plausibility function is

$$plaus(A) = 1 - bel(\overline{A})$$

It's the "plausibility" of A in that it's the degree to which the evidence fails to support its negation.

Partial Representation Theorem (Dempster, 1967)

Every belief function *bel* is a lower probability function, with *plaus* its corresponding upper probability function.

- Some lower probability functions are not belief functions.
  - Lower probabilities don't always satisfy (B3)

Shafer's Taxonomy of Belief Functions

- Belief Functions
  - Support Functions
    - Separable
      - Simple
      - Not Simple
    - Not Separable
  - Quasi Support Functions
    - Non-Bayesian
    - Bayesian (Nontrivial)

Dynamics

Dempster's Rule

Definition: *Dempster Combination*

If m_1 and m_2 are mass functions, their combination is denoted $m_1 \oplus m_2$ and is defined $m_1 \oplus m_2(\varnothing) = 0$, and for non-empty A:

$$(m_1 \oplus m_2)(A) \quad = \quad c \sum_{B,C:B \cap C = A} m_1(B) m_2(C)$$

where c is a normalizing constant.

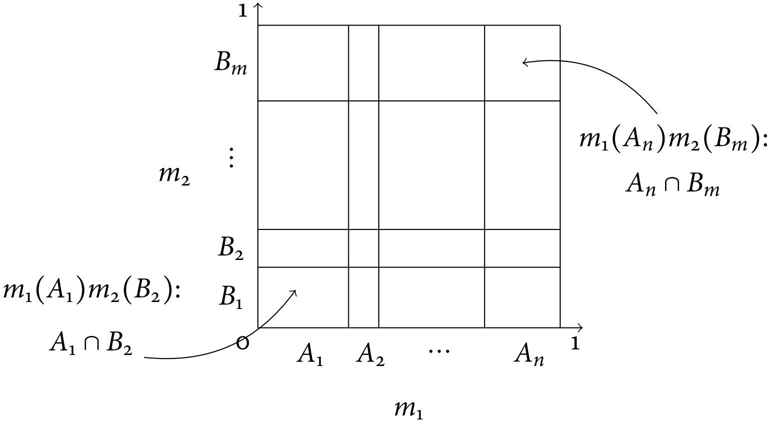
The normalizing constant is necessary to account for "leaks":

- Sometimes $B \cap C = \varnothing$ but $m_1(B) m_2(C) > 0$.
- Because of (M1), this mass must be thrown out.
- So we have

$$c \quad = \quad \left(1 - \sum_{B,C:B \cap C = \varnothing} m_1(B) m_2(C)\right)^{-1}$$

Visualizing Dempster's Rule

Dempster's rule is hard to grasp intuitively, but Shafer provides a helpful visualization:



Some Basics

- $m_1 \oplus m_2$ is undefined when c's denominator is 0.
- $m_1 \oplus m_2$ is always a mass function.
- We write $bel_1 \oplus bel_2$ for $m_1 \oplus m_2$'s belief function.
- Combination is associative and commutative:

$$\begin{aligned} bel_1 \oplus (bel_2 \oplus bel_3) &= (bel_1 \oplus bel_2) \oplus bel_3 \\ bel_1 \oplus bel_2 &= bel_2 \oplus bel_1 \end{aligned}$$

- Vacuous combination has no effect:

$$bel \oplus bel_0 = bel$$

- If $bel = bel_1 \oplus bel_2$ with corresponding commonality functions Q, Q_1, Q_2, then

$$Q(A) = c\, Q_1(A) Q_2(A)$$

Dempster Conditioning

Definition: Dempster Conditioning

Let bel_E be the belief function corresponding to the mass function assigning $m(E) = 1$. Then

$$bel(A|E) \quad =_{df} \quad bel \oplus bel_E$$

Theorem

$$\begin{aligned} bel(A|E) &= \frac{bel(A \cup \overline{E}) - bel(\overline{E})}{1 - bel(\overline{E})} \\ plaus(A|E) &= \frac{plaus(A \cap E)}{plaus(E)} \end{aligned}$$

Simple Support

A quick definition:

Definition: Simple Support Function

A belief function is a simple support function iff

$$bel(A) = \begin{cases} 0 & \text{if } S \nsubseteq A \\ s & \text{if } S \subseteq A \text{ but } A \neq \Omega \\ 1 & \text{if } A = \Omega \end{cases}$$

We say that *bel* is *focused on S*.

- Clearly, the corresponding mass function is

$$m(A) = \begin{cases} s & \text{if } A = S \\ 1 - s & \text{if } A = \Omega \\ 0 & \text{otherwise} \end{cases}$$

Special Case: Homogeneous Support

Consider two simple support functions both focused on A, with support degrees s_1 and s_2:

	A	Ω
A	$A \cap A = A$	$A \cap \Omega = A$
Ω	$\Omega \cap A = A$	$\Omega \cap \Omega = \Omega$

	$m_2(A) = s_2$	$m_2(\Omega) = 1 - s_2$
$m_1(A) = s_1$	$s_1 s_2$	$s_1(1 - s_2)$
$m_1(\Omega) = 1 - s_1$	$s_2(1 - s_1)$	$(1 - s_1)(1 - s_2)$

$$\begin{aligned}(m_1 \oplus m_2)(A) &= s_1 + s_2 - s_1 s_2 \\ (m_1 \oplus m_2)(\Omega) &= 1 - (m_1 \oplus m_2)(A)\end{aligned}$$

- Notice that $(m_1 \oplus m_2)(A) > s_1, s_2$
- Notice that $c = 1$.

Special Case: Heterogeneous Support

Two simple support functions focused on distinct A and B when $A \cap B \neq \varnothing$:

	B	Ω
A	$A \cap B$	A
Ω	B	Ω

	$m_2(B) = s_2$	$m_2(\Omega) = 1 - s_2$
$m_1(A) = s_1$	$s_1 s_2$	$s_1(1 - s_2)$
$m_1(\Omega) = 1 - s_1$	$s_2(1 - s_1)$	$(1 - s_1)(1 - s_2)$

$$
\begin{aligned}
(m_1 \oplus m_2)(A) &= s_1(1 - s_2) \\
(m_1 \oplus m_2)(B) &= s_2(1 - s_1) \\
(m_1 \oplus m_2)(A \cap B) &= s_1 s_2 \\
(m_1 \oplus m_2)(\Omega) &= (1 - s_1)(1 - s_2)
\end{aligned}
$$

- Again, $c = 1$.

Special Case: Conflicting Support

Two simple support functions focused on distinct A and B when $A \cap B = \varnothing$. Now $c = (1 - s_1 s_2)$.

	B	Ω
A	$\varnothing$	A
Ω	B	Ω

	$m_2(B) = s_2$	$m_2(\Omega) = 1 - s_2$
$m_1(A) = s_1$	$s_1 s_2$	$s_1(1 - s_2)$
$m_1(\Omega) = 1 - s_1$	$s_2(1 - s_1)$	$(1 - s_1)(1 - s_2)$

$$
\begin{aligned}
(m_1 \oplus m_2)(A) &= s_1(1 - s_2)/(1 - s_1 s_2) \\
(m_1 \oplus m_2)(B) &= s_2(1 - s_1)/(1 - s_1 s_2) \\
(m_1 \oplus m_2)(\Omega) &= (1 - s_1)(1 - s_2)/(1 - s_1 s_2)
\end{aligned}
$$

- Notice that $(m_1 \oplus m_2)(A) < m_1(A)$, and similarly for B.

Separable Support Functions

Definition: Separable Support Function

A belief function is separable iff it can be obtained by combining one or more simple support functions.

Separability Theorem (Shafer 1976: 90)

If *bel* is a non-vacuous, separable support function, there exists a unique collection of non-vacuous, simple support functions $bel_1, \ldots, bel_n$ such that

(1) $bel = bel_1 \oplus \ldots \oplus bel_n$

(2) The focus of each bel_i, S_i, is such that $bel(S_i) > 0$

(3) bel_i and bel_j have different foci when $i \neq j$.

- Note: separability does not assure us that *bel*'s actual history can be recovered; witness condition (3).
  - Recall the results of homogeneous combination.

Dempster Meets Jeffrey

Partial Representation Theorem (Shafer, 1981a)

Every Jeffrey update can be represented as a Dempster update. Whenever two probability functions are related by

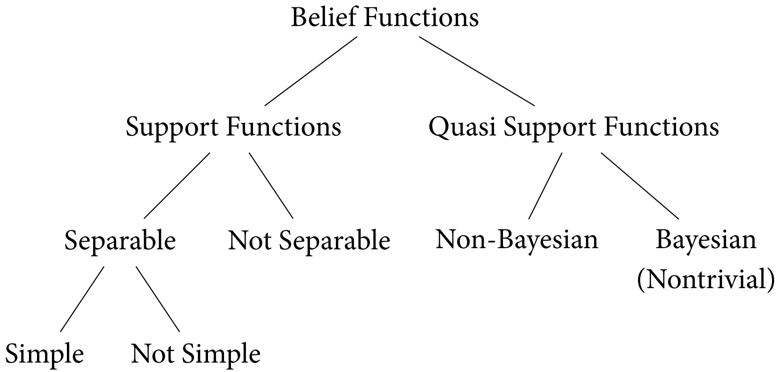
$$q(\cdot) = \sum_i p(\cdot|E_i)q(E_i)$$

for a partition $\{E_i\}$, there is a belief function *bel* such that $q = p \oplus bel$.

- *bel* will not be unique, generally speaking.
- *bel*'s focal elements will be unions of the E_i.
- Shafer (1981b) argues that the Dempster representation has the advantage of representing the evidence on its own, before prior belief is factored in. (Cf. (Field, 1978; Garber, 1980; Christensen, 1992; Lange, 2000; Wagner, 2002).)

Taxonomy

Reminder: Taxonomy of Belief Functions



Refinements

We've characterized the simple and separable support functions. What about the remaining support functions?

- Here we need the notion of a refinement; the division of a space's atoms into sub-possibilities.

Definition: *Refinement*

A map $r : 2^{\Omega} \to 2^{\Theta}$ is a refinement iff

(1) $r(\{w\}) \neq \varnothing$ for all $w \in \Omega$

(2) $r(\{w\}) \cap r(\{w'\}) = \varnothing$ if $w \neq w'$

(3) $\bigcup_{w \in \Omega} r(\{w\}) = \Theta$

(4) $r(A) = \bigcup_{w \in A} r(\{w\})$

Intuitively, r takes Ω's atoms to a nontrivial partition (1–3), and any larger set to the union of the sets corresponding to its atoms (4).

Restrictions

We can now characterize the restriction of a belief function over a refined space to the space from which it was refined.

Definition: *Restriction*

Let $r : 2^\Omega \rightarrow 2^\Theta$ be a refinement, and *bel* a belief function defined over Θ. The restriction of *bel* to Ω is written $bel|2^\Omega$, and is defined

$$bel|2^\Omega(A) = bel(r(A))$$

Theorem (Shafer, 1976: 126)

The restriction of a belief function is always a belief function.

Support Functions

Definition: Support Function

A belief function *bel* over Ω is a support function iff there is a refinement of Ω into Θ and a separable support function on 2^{Θ}, bel', such that $bel|2^{\Omega} = bel'$.

- Clearly, separable support functions are support functions.
- But some (even basic) support functions aren't separable.

Example: A Non-Separable Support Function

Suppose $\Omega = \{w_1, w_2, w_3\}$ and

$$m(\{w_1, w_2\}) = (\{w_2, w_3\}) = m(\Omega) = 1/3$$

Then bel_m is a support function, but is not separable.

Theorems 7.1 and 7.2 of Shafer (1976: 143) verify this.

Weighing Evidence

Given three natural assumptions, we can construct a nice measure of evidential weight.

1. Existence: the value of a simple support function focused on A is determined by a *weight of evidence* for A, w.
   - There is some function such that $g(w) = s$.
2. Scale: weights of evidence vary from 0 to ∞.
   - $g : [0, \infty] \to [0, 1]$.
3. Additivity: given two simple support functions focused on A, their combination is determined by the sum of their respective weights.
   - $g(w_1 + w_2) = (bel_1 \oplus bel_2)(A)$.

Weighing Evidence

Theorem (Shafer 1976: 77-8)

If g satisfies the following:

$g : [0, \infty] \to [0, 1]$

If $g(w_1) = s_2$, $g(w_2) = s_2$, then $g(w_1 + w_2) = s_1 + s_2 - s_1 s_2$

then $g(w) = 1 - e^{cw}$ for any constant c.

Choosing $c = 1$ for convenience, we measure weight by

$$g(w) \quad = \quad 1 - e^{-w}$$

Weighing Conflict

Two simple support functions conflict to the extent that they assign mass to incompatible propositions.

- The more mass thrown away

$$\kappa = \sum_{B,C:B\cap C=\varnothing} m_1(B)m_2(C)$$

  the greater the weight of conflict.

Definition: *Weight of Conflict*

$$\mathrm{Con}(bel_1, bel_2) = \log\left(\frac{1}{1-\kappa}\right)$$

The Weight-of-Conflict Conjecture

Definition: *Weight of Internal Conflict*

If bel is a separable support function, it's weight of internal conflict, W_{bel}, is $\text{Con}(bel_1, \ldots, bel_n)$, where $bel_1 \oplus \cdots \oplus bel_n$ is bel's canonical decomposition into simple support functions.

Conjecture (Shafer 1976: 96)

Let bel_1 and bel_2 be separable support functions with commonality functions Q_1, Q_2, and weights of internal conflict W_1, W_2. Then, if $Q_1(A) \leq Q_2(A)$ for all A, $W_1 \geq W_2$.

Limits of Sequences of Belief Functions

Recall how to take the limit of a sequence of functions:

Definition: Limit of a Sequence of Functions

Suppose $f_1, f_2, \ldots$ is an infinite sequence of functions. Then its limit is f iff

$$\lim_{i \to \infty} f_i(A) \quad = \quad f(A)$$

for all A in the domain.

Then we have the following theorem about belief functions:

Theorem (Shafer 1976: 200)

If a sequence of belief functions has a limit, the limit is a belief function.

Quasi Support Functions

We're finally in a position to characterize the remaining belief functions, the quasi support functions:

- They are the limits of sequences of separable support functions over a more refined space.

Theorem (Shafer 1976: 200)

If *bel* is not a support function, it is the restriction of a limit of a sequence of separable support functions.

That is, there is a refinement of Ω into Θ and a sequence of separable support functions $bel_1, bel_2, \ldots$ on Θ such that

$$bel = \left(\lim_{i \to \infty} bel_i\right) |2^{\Omega}$$

Two Notes for Later

Two things to note for use in a moment:

1. Given that the bel_i are separable,

$$\left(\lim_{i\to\infty} bel_i\right)|2^{\Omega} = \lim_{i\to\infty}\left(bel_i|2^{\Omega}\right)$$

2. Each $bel_i|2^{\Omega}$ is a support function.

So we can also say that the above *bel* is the limit of a sequence of support functions.

Some Examples

As it turns out, all non-trivial probability functions are quasi support functions.

Theorem (Shafer 1976: 201)

If *bel* is a belief function with at least one $A \subseteq \Omega$ such that $bel(A) > 0$ and $bel(A) + bel(\overline{A}) = 1$, then *bel* is a quasi support function.

But other examples abound, even very elementary ones.

Example: A Non-probabilistic Quasi Support Function

Let $\Omega = \{w_1, w_2, w_3\}$ and $m(\{w_1, w_2\}) = m(\{w_2, w_3\}) = 1/2$. Then bel_m is not a support function, i.e. it is a quasi support function.

- Follows from Shafer's Theorem 7.1; again, I'm not sure whether there is a more direct way to see this.

Weights of Impinging Evidence

Definition: Impingement Function

If *bel* is a separable support function with w its corresponding weight-of-evidence function, its impingement function is defined

$$V(A) = \sum_{B : A \notin B} w(B)$$

V is the weight of evidence for propositions compatible with $\overline{A}$.

- Weights of evidence are additive, by assumption.

Intuitively, $V(A)$ is the weight of evidence impinging on A.

- Each $w(B)$ "impugns" part of A, since $A \notin B$.

Infinite, Contradictory Evidence

Theorem (Shafer 1976: 220 –1)

If $bel_1, bel_2, \ldots$ is a sequence of separable support functions whose limit is not a separable support function, and $V_1, V_2, \ldots$ are the corresponding impingement functions, then

$$\lim_{i \to \infty} V_i(\{w\}) \quad = \quad \infty$$

for every $w \in \Omega$.

> *"Because of the dubious nature of such infinite contradictory weights of evidence, it is natural to call a belief function a quasi support function whenever it is not a support function but is the limit of a sequence of separable support functions or the restriction of such a limit." (Shafer 1976: 201)*

Two Worries

Is this interpretation of the results too quick? Two reasons I'm suspicious:

1. We haven't shown that quasi support functions are the limits of sequences of separable support functions, only that they are the *restrictions of* such limits.
    - Some quasi support functions are limits of sequences of separable support functions, not merely restrictions of such limits.
    - But some are only "indirectly" so, i.e. restrictions of such limits.
        - (Or, using our earlier two notes, limits of sequences of support functions, though not necessarily separable ones.)

   How does the theorem tell us that quasi support functions obtainable only as restrictions represent "infinite contradictory weights of evidence"?

Two Worries (continued)

Is this interpretation of the results too quick? Two reasons I'm suspicious:

2. But suppose we focus just on those quasi support functions that are directly limits of separable support functions.

   That the weights supporting contradictory propositions tend to infinity does not obviously entail that the function at the limit itself represents such evidence.

   - For one thing, these are the limits of infinite *sequences*, not infinite *combinations*.
   - For another, they are limits at *infinity*, and the finite-transfinite gap is notoriously tricky.
     - Examples: Adam & Eve, Infinity Bank™

Infinite Evidence & Statistics

Barring these concerns, the result is quite striking:

- Bayesian rationality demands that we believe as if we had infinite evidence for contradictory propositions!

How can this be?

> *"Those who are accustomed to thinking of partial beliefs based on chances as paradigmatic may be startled to see them relegated to a peripheral role and classified among those partial beliefs that cannot arise from actual, finite evidence. But students of statistical inference are quite familiar with the conclusion that a chance cannot be evaluated with less than infinite evidence." (Shafer 1976: 201)*

Yea Ok, But *Contradictory*?!

What about the fact that such hypothetical evidence is not just infinite, but supporting of *contradictory* propositions?

> *"To establish a value between zero and one as the chance for a given outcome of an aleatory process, one must obtain the results of an infinite sequence of independent trials of the process [...] One could ask for no better example of infinite, precisely balanced and unobtainable evidence." (Shafer 1976: 201-2)*

In other words: if we had had enough evidence to determine the true chances for the next flip of a coin, we would have evidence of infinite weight that the next flip will be heads.

- Notice a corollary: it is possible to have evidence of infinite weight supporting *no* confidence.

Decisions

Three Approaches

There are at least two broad approaches to constructing a decision rule for DST:

1. Fall back on the theory of upper and lower probabilities.
2. Collapse the belief function into a probability function.
3. Make assumptions justified by specifics of the application.

Upper & Lower Expectations

Belief functions can be seen as inducing expectation intervals:

- For every belief function *bel*, there is a canonical set of probability functions $\mathcal{P}$ such that $bel = \mathcal{P}_*$ and $plaus = \mathcal{P}^*$:

$$\mathcal{P} = \{p : p(A) \geq bel(A) \text{ for all } A\}$$

- So we can define $E_{bel} = \underline{E}_{\mathcal{P}}$ and $E_{plaus} = \overline{E}_{\mathcal{P}}$.
- We can then fall back on rules like Total Domination.

An important caveat:

- Generally, several $\mathcal{P}$s can be associated with a given *bel*.
- Some decision rules, like Levi's, depend not only on the interval $[\mathcal{P}_*, \mathcal{P}^*]$, but on the particular contents of $\mathcal{P}$.
- For such decision rules, which $\mathcal{P}$ we associate with *bel* matters, so a canonical translation is required.

Cutting Out the Middle Man

Can we avoid the detour through upper/lower probabilities?

- Shafer and many others explicitly reject the upper/lower probability interpretation of belief functions.
- It'd be computationally easier to cut out the middle man.

Answer: yes!

- Recall that probabilistic expectation can be re-expressed

$$\begin{aligned} E_p(X) &= \sum_i^n p(X = x_i)x_i \\ &= x_1 + \sum_{i=1}^{n-1} p(X > x_i)(x_{i+1} - x_i) \end{aligned}$$

Cutting Out the Middle Man (continued)

We can defined DST's expected values in the same way:

Definition: E_{bel} and E_{plaus}

$$\begin{aligned} E_{bel}(X) &= x_1 + \sum_{i=1}^{n-1} bel(X > x_i)(x_{i+1} - x_i) \\ E_{plaus}(X) &= x_1 + \sum_{i=1}^{n-1} plaus(X > x_i)(x_{i+1} - x_i) \end{aligned}$$

Theorem (Schmeidler 1986)

If $\mathcal{P}$ is the canonical set of probability functions associated with *bel*, then $E_{bel} = \underline{E}_{\mathcal{P}}$ and $E_{plaus} = \overline{E}_{\mathcal{P}}$.

The Transferable Belief Model

Smets and Kennes (1994) proposed the TBM, which distinguishes two levels of degree of belief:

- Credal: obeys rules of of DST.
- Pignistic: obeys rules of probability.

When a decision must be made, we "flatten" the mass function into a probability function, and use good ol' expected utility.

Definition: Pignistic Probability (TBM)

Given a mass function m, the pignistic probability function corresponding to m, p_m, is defined:

$$p_m(\{w\}) = \sum_{A : w \in A} \frac{m(A)}{|A|}$$

for all $w \in \Omega$, where $|A|$ is the cardinality of A.

Worries About TBM

Two worries about the TBM:

1. Frame Dependence: the pignistic probabilities obtained from a given mass function depend heavily on the granularity of Ω.
   - Mass is distributed according to the size of A.
   - So a finer division of A means a higher $p_m(A)$.
   - So the problems with the principle of indifference effectivey return in the decision theoretic context.
2. Dutch Books: pignistic probabilities are dynamically Dutch bookable.
   - The dynamics of pignistic probabilities do not obey conditionalization.
   - Smets (1994) insists that dynamic Dutch books don't arise because of the distinction between "hypothetical" facts and "factual" facts. (?!?)
   - See Snow (1998) for a rebuttal.

Making Assumptions

Strat (1994) proposes associating a parameter ρ with an "uncommitted" mass assignment.

- ρ varies from 1 to 0 according as we think nature will resolve the "unknown" probability "favourably".
- Simplest case: m assigns all its values to atoms but one.

$$E_\rho(X) \quad = \quad E_{bel}(X) + \rho[E_{plaus}(X) - E_{bel}(X)]$$

- ρ is reminiscent of, and inspired by, Hurwicz's (1952) optimism index.
- Lesh (1986) makes a similar proposal; Strat views Lesh's as differring in two respects:
  - Lesh's parameter reflects empirical commitments.
  - Lesh's parameter is used for a linear interpolation of the range of possible probabilities; Strat's for the range of expected values.

Other Topics: Independence

Defining independence in DST is trickier than for probability theory.

- The usual definition, $bel(A|B) = bel(A)$, doesn't work.
- Several other definitions have been proposed.
- Question: how do they interact with updating?
  - Probabilistic independence on the evidence is preserved by conditioning rules.
  - I think this is deeply problematic for Bayesianism. (Weisberg 2009, manuscript)
  - Does something analogous hold of Dempster's rule? See (Ben Yaghlane, Smets and Mellouli 2000, 2002) for some discussion.

Other Topics: Interpretation

How should we understand 'degree of belief' in DST?

- Bayesians provide many heuristics and operationalizations to help us get a grip on the notion of credence.
- Shafer (1981a,b) and Shafer and Tversky (1983, 1985) offer a heuristic where chancy translation is the canonical scale.
- Pearl (1988) argues for an interpretation in terms of probability of provability.
- See (Smets 1994) for a survey of some standard interpretations.

References I

[1] Boutheina Ben Yaghlane, Philipe Smets, and Khaled Mellouli.
Independence concepts for belief functions.
In *Proceedings 8th International Conference IPMU Information Processing and Management of Uncertainty in Knowledge-based Systems*, volume 1, pages 357–364, 2000.

[2] Boutheina Ben Yaghlane, Philipe Smets, and Khaled Mellouli.
Belief function independence: I. the marginal case.
International Journal of Approximate Reasoning, 29(1):47–70, 2002.

[3] David Christensen.
Confirmational holism and bayesian epistemology.
Philosophy of Science, 59, 1992.

[4] Arthur P. Dempster.
Upper and lower probabilities induced by a multivalued mapping.
Annals of Mathematical Statistics, 38(2):325–339, 1967.

[5] Hartry Field.
A note on jeffrey conditionalization.
Philosophy of Science, 45, 1978.

[6] Daniel Garber.
Field and jeffrey conditionalization.
Philosophy of Science, 47, 1980.

[7] Marc Lange.
Is jeffrey conditionalization defective by virtue of being non-commutative? remarks on the sameness of sensory experience.
Synthese, 123(3):393–403, 2000.

References II

[8] Stephen A. Lesh.
An Evidential Theory Approach to Judgment-Based Decision Making.
Phd thesis: Department of forestry and environmental studies, Duke University, 1986.

[9] Judea Pearl.
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann Publishers Inc., 1988.

[10] D. Schmeidler.
Integral representation without uncertainty.
Proceedings of the American Mathematical Society, 97(2):255–261, 1986.

[11] Glenn Shafer.
A Mathematical Theory of Evidence.
Princeton University Press, 1976.

[12] Glenn Shafer.
Constructive probability.
Synthese, 48(1):1–60, 1981.

[13] Glenn Shafer.
Jeffrey's rule of conditioning.
Philosophy of Science, 48(3):337–362, 1981.

[14] Glenn Shafer and Amos Tversky.
Weighing evidence: The design and comparison of probability thought experiments.
Technical report, Department of Psychology, Stanford University, 1983.

[15] Glenn Shafer and Amos Tversky.
Languages and designs for probability judgment.
Cognitive Science, 9(3):309–339, 1985.

References III

[16] Philipe Smets.
What is dempster-shafer's model.
In Ronald R. Yager, editor, *Advances in the Dempster-Shafer Theory of Evidence*. John Wiley and Sons, Inc., 1994.

[17] Philipe Smets and Robert Kennes.
The transferable belief model.
Artificial Intelligence, 66(2):191–234, 1994.

[18] Paul Snow.
The vulnerability of the transferable belief model to dutch books.
Artificial Intelligence, 105(2):345–354, 1998.

[19] Thomas M. Strat.
Decision analysis using belief functions.
In Ronald R. Yager, editor, *Advances in the Dempster-Shafer Theory of Evidence*. John Wiley and Sons, Inc., 1994.

[20] Carl Wagner.
Probability kinematics and commutativity.
Philosophy of Science, 69:266–278, 2002.

[21] Jonathan Weisberg.
Commutativity or holism: A dilemma for conditionalizers.
British Journal for the Philosophy of Science, 60(4):793–812, 2009.

[22] Jonathan Weisberg.
Updating, undermining, and independence.
manuscript.