4 Dempster-Shafer Theory

4.1 Introduction

Dempster-Shafer theory (DS theory), also known as *evidence theory* or *theory of belief functions*, is a mathematical theory of evidence and plausible reasoning. The theory has been developed by Glenn Shafer based on earlier work of Arthur Dempster. Already in the early eighties the theory received much attention as a promising theory for the handling of uncertain information in expert systems.

The formalism of DS has several interpretations. Dempster's model, in which belief functions are defined by means of multivalued mappings, is treated in the chapter on generalised probability theory. In this chapter, we concentrate on Shafer's interpretation of the formalism given in his book *A Mathematical Theory of Evidence*. This book contains most of the material of this chapter (and much more).

As a theory of evidence, DS theory has some advantages over probability theory. Important advantages are the ease with which evidence of different levels of abstraction can be represented and the possibility of discriminating between uncertainty and ignorance. However, the main attraction of DS theory is the availability of a rule to combine the effect of different bodies of evidence, namely Dempster's combination rule.

The lack of a combination rule in probability theory is an important source of the combinatorial problems one encountered when applying probability theory to reasoning with uncertainty in knowledge-based systems. (Cf. chapter 1.)

Dempster's rule is in fact the main instrument of DS theory. But the justification of this rule turns out to be problematic, which makes the whole theory suspect from a theoretical point of view. Still, DS theory is considered to be a serious alternative to probability theory by quite a few researchers in both theoretical and applied AI.

After introducing the terminology of DS theory, Dempster's rule is discussed at length.

4.2 Basic Terminology

DS theory has its own (extensive) terminology, partly because DS theory contains many new notions, partly because well-known notions are given a new name. In the context of DS theory, a sample space is called a *frame of discernment*, or simply *frame*. As before, we assume the frames to be finite.

Partial belief concerning the events of a frame Ω is represented by a belief function over Ω. As before, belief functions are ∞-monotone capacities, but in DS theory they are given an intuitively appealing interpretation via mass functions.

Definition 4.2.1 *A* mass function*, or* basic probability assignment (bpa)*, over a frame Ω is a function $m : 2^{\Omega} \to [0,1]$ such that the following two conditions hold.*

$$m(\emptyset) = 0. \tag{60}$$

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{61}$$

The quantity $m(A)$ is a measure of the belief that is assigned to *exactly* the set A (and not to any proper subset of A). If $m(A) = 0.4$, then a 40% portion of one's total belief is assigned to exactly A. Equation (61) formalises the convention that one's total belief has measure one, and equation (60) reflects the requirement that none of the belief of a rational agent is assigned to the impossible event.

A measure of the (total) belief in A should also take into account the measures of belief assigned to more specific propositions, i.e., to subsets of A. For this purpose belief functions induced by mass functions are introduced.

Definition 4.2.2 *Let m be a mass function over a frame Ω. The* belief function *Bel* induced by *m is defined as follows.*

$$\text{For every } A \subseteq \Omega,\ Bel(A) = \sum_{B \subseteq A} m(B). \tag{62}$$

Example 4.2.1 (The Safecracker) Investigating the theft of some important documents from a safe, Sherlock Holmes comes up with the following two clues.

1. Examination of the safe suggests, with a high degree of certainty, say 70%, that the safecracker was left-handed.
2. Since the door giving entrance to the room with the safe has not been forced, it can be concluded, again with a high degree of certainty, say 80%, that it was an "inside job".

Consider the first clue, and notice that it gives no support whatsoever to the hypothesis that the safecracker was right-handed, it does not imply with absolute certainty that the safecracker was left-handed, and it does not point to any particular left-handed individual. The representation of evidence in DS theory does justice to all these points.

Let $\Omega = \{x : x \text{ is a possible suspect}\}$ be the frame consisting of the possible safecrackers. (We assume that exactly one of them is the actual safecracker.) Let $L = \{x \in \Omega : x \text{ is left-handed}\}$. Then the first clue is represented by the mass function m_1 over Ω given by $m_1(L) = 0.7$ and $m_1(\Omega) = 0.3$. (It automatically follows that zero mass is assigned to sets other than L and Ω.)

The belief function Bel_1 induced by m_1 is given as follows.

$$Bel_1(A) = \begin{cases} 1 & \text{if } A = \Omega \\ 0.7 & \text{if } L \subseteq A \neq \Omega \\ 0 & \text{otherwise.} \end{cases}$$

Thus the belief based on the first clue in the proposition that the safecracker was left-handed measures 0.7 on a scale of 0 to 1, whereas both the proposition that the safecracker was right-handed and the proposition that John "Lefty" Jones is the safecracker are assigned a zero measure of belief.

The class of belief functions induced by mass functions coincides with the previously introduced class of belief functions as ∞-monotone capacities.

Proposition 4.2.1 *Bel is a belief function over Ω iff Bel is the belief function induced by some mass function m over Ω.*

Proposition 4.2.2 *If Bel is a belief function over Ω, then there is a unique mass function m over Ω such that Bel is the belief function induced by m. This mass function is given by the following equation.*

$$\text{For all } A \subseteq \Omega,\ m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \cdot Bel(B). \tag{63}$$

$Bel(A)$ is a measure of the (total) belief that is certainly assigned to A. There may also be belief that could possibly be assigned to A, namely belief assigned to some B which is consistent with A. Plausibility functions take also this belief into account.

Definition 4.2.3 *Let m be a mass function over a frame Ω. The* plausibility function Pl induced by m *is defined as follows.*

$$\text{For every } A \subseteq \Omega,\ Pl(A) = \sum_{A \cap B \neq \emptyset} m(B). \tag{64}$$

$Pl(A)$ is a measure of the belief which is not (yet) assigned to propositions which imply the falsity of A. It is easy to check that Pl and Bel are dual. Since, by proposition 4.2.2, it is also possible to recover from Bel its unique underlying mass function m, the above defined functions m, Bel, Pl are interdefinable.

To this list of interdefinable functions one can add yet another function, namely the commonality function, defined as follows.

Definition 4.2.4 *Let m be a mass function over a frame Ω. The* commonality function Q induced by m *is defined as follows.*

$$\text{For every } A \subseteq \Omega,\ Q(A) = \sum_{A \subseteq B \subseteq \Omega} m(B). \tag{65}$$

$Q(A)$ can be understood as a measure of the belief which can move to every element of A. Commonality functions are sometimes used to simplify the expression and verification of properties of belief functions.

We introduce some further notions. A subset A of a frame Ω is called a *focal element* of a belief function Bel over Ω iff $m(A) > 0$. (Here it is assumed to be understood that the m denotes the mass function associated with Bel. Similar assumptions will be made frequently.) The union of all focal elements of a belief function is called its *core*.

The complete information about the measure of belief in A can be represented by the *belief interval* $[Bel(A), Pl(A)]$, where $Pl(A) - Bel(A)$ is a natural expression of the ignorance concerning A. Complete ignorance with respect to a frame Ω is represented by the *vacuous belief function* over Ω, which is the belief function induced by the mass function m defined by $m(\Omega) = 1$ and for all $A \subset \Omega, m(A) = 0$.

For any subset A of a frame Ω, we write 1_A for the mass function which assigns mass 1 to the set A (and zero mass to all other events). We will use the same notation for the associated belief function, plausibility function, et cetera, and let the context determine which function is meant. Thus, the vacuous belief function over Ω can be denoted by 1_Ω.

Probability functions over a frame are a special kind of belief functions.

Definition 4.2.5 *A belief function Bel over a frame Ω is called a* Bayesian belief function *iff Bel is a probability function over Ω.*

The following proposition lists several characterisations of Bayesian belief functions.

Proposition 4.2.3 *Let Bel be a belief function over a frame Ω. The following are equivalent.*

1. *Bel is Bayesian.*
2. *$Bel = Pl$.*
3. *All the focal elements of Bel are singletons.*
4. *For every $A \subseteq \Omega$, if $Q(A) > 0$, then A is a singleton.*
5. *For every $A \subseteq \Omega$, $Bel(A) + Bel(\overline{A}) = 1$.*

It is tempting to consider $Bel(A)$, respectively $Pl(A)$, as lower, respectively upper, bound of the "true" probability of A. Although this interpretation is in accordance with the original intention of Arthur Dempster, it is explicitly rejected by Glenn Shafer. In his view, DS theory is pre-eminently suited for situations in which there is not sufficient evidence available to speak of the true probabilities of all relevant propositions.

In Dempster's model, as described in chapter 3, the belief and plausibility functions are explicitly linked to a probability function over a frame (the original information level) which is related to the frame representing the target information level. In this chapter, the belief functions, mass functions, et cetera, have been introduced without referring to probability theory, just as in Shafer's book. Only the target information level is mentioned, the original information level is not even alluded to.

It should be mentioned that in later publications Shafer comes quite close to Dempster's model. Presently, the *Transferable Belief Model* of Philippe Smets is probably the best developed theory which takes Shafer's original interpretation serious.

To end this section, we show how mass functions can be introduced in Dempster's model, and we prove that not every belief function over Ω is an inner measure extension over Ω. (This fact was already mentioned, but not proved, in chapter 3).

Definition 4.2.6 *Let* $\langle \Theta, 2^\Theta, P \rangle$ *be a probability space,* Ω *is a sample space, and* $\Gamma : \Theta \to 2^\Omega \backslash \{\emptyset\}$*. The mass function* m_Γ *is defined as follows.*

$$\textit{For every } A \subseteq \Omega, m_\Gamma(A) = P(\{\theta \in \Theta : \Gamma(\theta) = A\}). \tag{66}$$

Proposition 4.2.4 *Let Bel be a belief function over* Ω*. If Bel is an inner measure extension, then its focal elements are pairwise disjoint.*

Proof. Assume that $Bel = P_*$, for some $P \in PROB(\Sigma, \Omega)$. Let Θ be a basis of Σ. Then $Bel = P'_{\Gamma,low}$, where Γ is id_Θ and $P' \in PROB(\Theta)$ is defined by $P'(\{\theta\}) = P(\theta)$. (Cf. proposition 3.2.3.) The focal elements of Bel are the elements of Θ, which are pairwise disjoint. ∎

It follows that the belief function Bel_1 of example 4.2.1 is not an inner measure extension.

Exercise 4.2.1 Give the plausibility function Pl_1 induced by m_1 of example 4.2.1.

Exercise 4.2.2 Let m be a mass function over Ω and let Bel and Pl be its associated belief and plausibility function. Show that Bel and Pl are dual functions.

Exercise 4.2.3 Formalise the second clue of example 4.2.1, by giving both the relevant mass function and the induced belief function.

Exercise 4.2.4 Let Pl be a plausibility function over Ω. Show that for any $A, B \subseteq \Omega, Pl(A \cup B) \leq Pl(A) + Pl(B)$.

Exercise 4.2.5 Give an example which shows that the following inequality does *not* hold.

$$Bel(A \cup B) \leq Bel(A) + Bel(B).$$

Exercise 4.2.6 Give an example which shows that the following inequality does *not* hold.

$$Bel(A \cup B) \geq Bel(A) + Bel(B).$$

Exercise 4.2.7 Prove proposition 4.2.3.

Exercise 4.2.8 Show that m_Γ of definition 4.2.6 is a mass function.

4.3 Dempster's Combination Rule

Consider example 4.2.1. There are *two* pieces of evidence which both can be represented by means of a belief function. In this section we discuss a rule, introduced by Dempster, for combining two belief functions into a new belief function which is supposed to represent the combined evidence.

Let us first discuss informally what the combined evidence in example 4.2.1 is. The first piece of evidence points (with degree 0.7) towards L, the set of left-handed individuals, whereas the second clue points (with degree 0.8) towards I, the set of "insiders". Taken together, the evidence should therefore support to some degree that the safecracker was a left-handed insider.

The support of L based on the first clue does not seem to be strengthened or weakened by the second clue. Thus, based on both clues, a degree 0.7 of belief in L seems reasonable. Similarly, the combined evidence suggests a degree 0.8 of belief in I. How much support does the combined evidence give to $L \cap I$? Surely less than 0.7 and 0.8. A reasonable answer might be 0.56 (the product of 0.7 and 0.8).

One can formulate the following rule. If the first piece of evidence is represented by a mass function m_1 which assigns mass $m_1(A)$ to A, and the second piece of evidence is represented by m_2 which assigns mass $m_2(B)$ to B, then the combined evidence assigns mass $m_1(A) \cdot m_2(B)$ to $A \cap B$.

Applying this rule to example 4.2.1 results in the following mass function m as a representation of the combined evidence. $m(L \cap I) = m_1(L) \cdot m_2(I) = 0.56$, $m(L) = m_1(L) \cdot m_2(\Omega) = 0.14$, $m(I) = m_1(\Omega) \cdot m_2(I) = 0.24$, and $m(\Omega) = m_1(\Omega) \cdot m_2(\Omega) = 0.06$. Since $Bel(L) = m(L) + m(L \cap I) = 0.7$, the belief in L is not changed. Similarly for the belief in I. New is the assignment of a degree 0.56 of belief in $L \cap I$.

To introduce Dempster's combination rule in its full generality, let m_1 and m_2 be the mass functions of the belief functions Bel_1 and Bel_2 with focal elements $A_1, \ldots, A_p$ and $B_1, \ldots, B_q$, respectively. The masses assigned by the mass functions can be visualised as segments of the unit interval, as shown in figure 5.

Figure 6 shows how the two intervals of figure 5 can be orthogonally combined to obtain a unit square representing the total mass assigned by $Bel_1 \oplus Bel_2$, the combination of Bel_1 and Bel_2, where Bel_1 commits vertical strips to its focal elements and Bel_2 horizontal ones.

Consider the intersection of the vertical strip representing the mass $m_1(A_i)$ assigned to A_i and the horizontal strip representing the mass $m_2(B_j)$

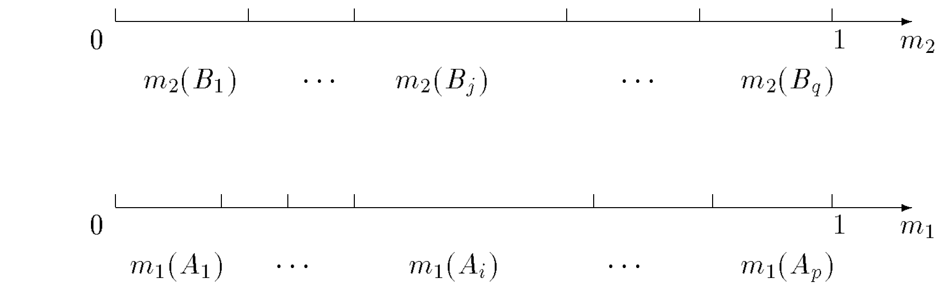

Figure 5: Masses assigned by m_1 and m_2.

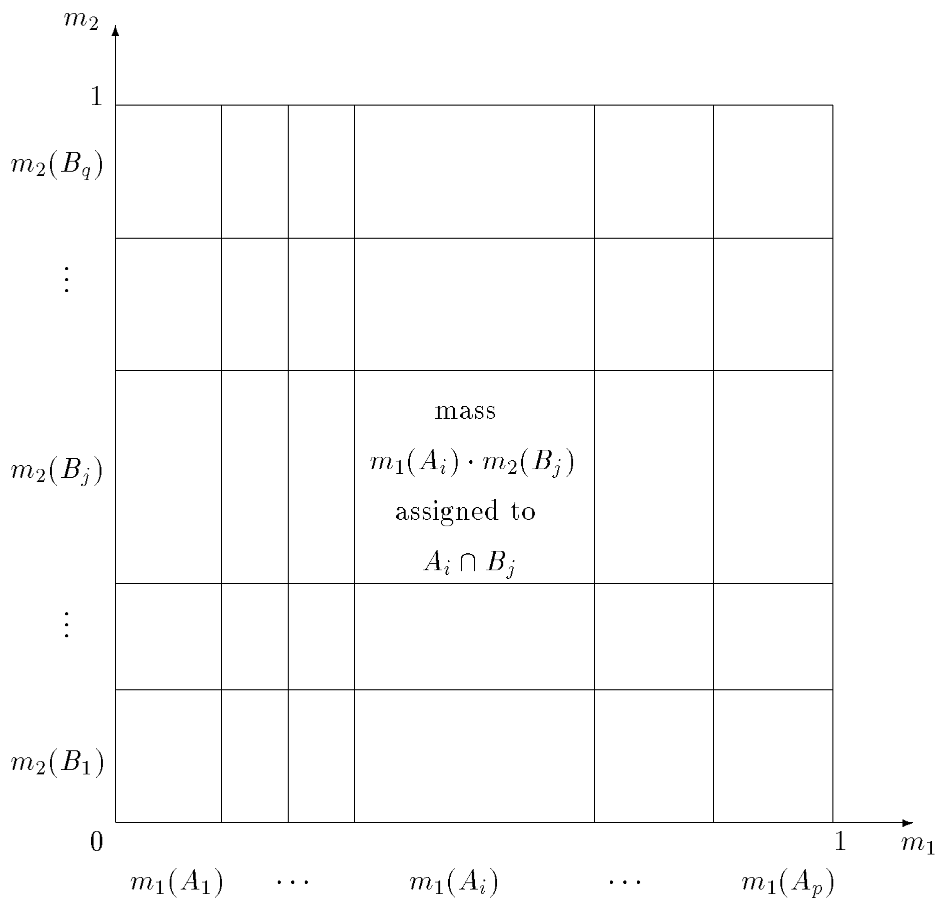

Figure 6: Mass assignment according to Dempster's rule.

assigned to B_j. This intersection represents mass assigned to both A_i and B_j, and thus to $A_i \cap B_j$. The mass represented by this intersection is assumed to measure $m_1(A_i) \cdot m_2(B_j)$, i.e., the area of the rectangle.

A given subset A of Ω may of course be the intersection of different pairs A_i and B_j. Hence to obtain the total mass assigned to (exactly) A by the combination of m_1 and m_2 (notation $m_1 \oplus m_2$) we have to take the sum (over i and j) of all products $m_1(A_i) \cdot m_2(B_j)$ such that $A = A_i \cap B_j$.

However, in this way some probability mass can be assigned to the empty set, since there can be a focal element A_i of m_1 and a focal element B_j of m_2 such that $A_i \cap B_j = \emptyset$. But then $m_1 \oplus m_2$ would fail to be a mass function.

To solve this problem, all rectangles representing mass assigned to the empty set are discarded and the measures of the remaining rectangles are rescaled by dividing through the sum of all $m_1(A_i) \cdot m_2(B_j)$ such that $A_i \cap B_j \neq \emptyset$, provided this sum does not equal 0; otherwise we say that Bel_1 and Bel_2 are not combinable. Hence we arrive at the following definition.

Definition 4.3.1 *Assume that Bel_1 and Bel_2 are belief functions over Ω induced by mass functions m_1 and m_2 such that $\sum_{A_i \cap B_j \neq \emptyset} m_1(A_i) \cdot m_2(B_j) \neq 0$. The* combination of Bel_1 and Bel_2 by Dempster's rule, *is the belief function $Bel_1 \oplus Bel_2$ induced by the mass function $m_1 \oplus m_2$ over Ω, defined as follows.*

$$m_1 \oplus m_2(A) = \begin{cases} 0 & \textit{if } A = \emptyset \\ \frac{\sum_{A_i \cap B_j = A} m_1(A_i) \cdot m_2(B_j)}{\sum_{A_i \cap B_j \neq \emptyset} m_1(A_i) \cdot m_2(B_j)} & \textit{if } \emptyset \neq A \subseteq \Omega. \end{cases} \tag{67}$$

The function $Bel_1 \oplus Bel_2$ is sometimes called the *orthogonal sum of Bel_1 and Bel_2*. The factor $[\sum_{A_i \cap B_j \neq \emptyset} m_1(A_i) \cdot m_2(B_j)]^{-1}$ is called the *renormalising constant* of Bel_1 and Bel_2. If $\sum_{A_i \cap B_j \neq \emptyset} m_1(A_i) \cdot m_2(B_j) = 0$, then $Bel_1 \oplus Bel_2$ is undefined, and Bel_1 and Bel_2 are called *not combinable.*

Proposition 4.3.1 *Let Bel_1 and Bel_2 be belief functions over Ω. The following are equivalent.*

1. *$Bel_1 \oplus Bel_2$ is undefined.*

2. *The cores of Bel_1 and Bel_2 are disjoint.*

3. *$Bel_1(A) = 1$ and $Bel_2(\overline{A}) = 1$, for some $A \subseteq \Omega$.*

4. *$Q_1(A) \cdot Q_2(A) = 0$, for all non-empty $A \subseteq \Omega$.*

The following proposition lists some properties of Dempster's combination rule. First we introduce a useful notation.

$$x \simeq y \overset{\text{def}}{\Longleftrightarrow} \begin{cases} x \text{ and } y \text{ are defined, and } x = y \\ \text{or} \\ x \text{ and } y \text{ are undefined} \end{cases}$$

Proposition 4.3.2 *Let Bel_1, Bel_2, and Bel_3 be belief functions over Ω.*

1. $Bel_1 \oplus Bel_2 \simeq Bel_2 \oplus Bel_1$.
2. $Bel_1 \oplus (Bel_2 \oplus Bel_3) \simeq (Bel_1 \oplus Bel_2) \oplus Bel_3$.
3. *If $Bel_1 \oplus Bel_2$ is defined and Bel_1 is Bayesian, then $Bel_1 \oplus Bel_2$ is also Bayesian.*
4. $Bel_1 \oplus 1_\Omega = Bel_1$.

It follows that the result of Dempster's combination rule does not depend on the order in which pieces of evidence are combined, since the rule is commutative (first property above) and associative (second property). The third property states that the combination of a belief function with a Bayesian belief function is again Bayesian (provided the belief functions are combinable). The fourth property states that combining with the vacuous belief function does not change one's degrees of belief.

To illustrate the use of commonality functions, we show how they simplify the formulation of Dempster's rule.

Proposition 4.3.3 *Assume that Bel_1 and Bel_2 are combinable belief functions over Ω, with commonality functions Q_1 and Q_2, respectively. Let k denote the renormalising constant of Bel_1 and Bel_2. The commonality function $Q_1 \oplus Q_2$ associated with the belief function $Bel_1 \oplus Bel_2$ is given as follows.*

$$Q_1 \oplus Q_2(A) = \begin{cases} 0 & \text{if } A = \emptyset \\ k \cdot Q_1(A) \cdot Q_2(A) & \text{if } \emptyset \neq A \subseteq \Omega. \end{cases} \tag{68}$$

An interesting special case of Dempster's rule is obtained when one of the belief functions to be combined is equal to 1_A, for some subset A of the frame. Since the function 1_A carries exactly the information that A is the case, combining a belief function Bel with 1_A is an intuitively appealing representation of conditioning Bel on A. Therefore, this special case of Dempster's combination rule is called *Dempster's rule of conditioning.*

Proposition 4.3.4 *Let Bel be a belief function over Ω such that $Pl(A) > 0$.*

$$\textit{For all } B \subseteq \Omega \text{ , } Bel \oplus 1_A(B) = \frac{Bel(B \cup \overline{A}) - Bel(\overline{A})}{1 - Bel(\overline{A})}. \tag{69}$$

$$\textit{For all } B \subseteq \Omega \text{ , } Pl \oplus 1_A(B) = \frac{Pl(B \cap A)}{Pl(A)}. \tag{70}$$

We sometimes write $Bel(\cdot|_{\oplus}A)$ or $Bel_{\oplus A}$ for the function $Bel \oplus 1_A$. Similar abbreviations are used for the plausibility function.

In chapter 3 we have seen that a belief function Bel over Ω can be viewed as (a special kind) of lower probability function and that Bel induces a set of probability functions $\langle Bel \rangle$ $(= \{P \in PROB(\Omega) : Bel \leq P\})$. Dempster's rule of conditioning and definition 3.3.2 of conditioning the set $\langle Bel \rangle$ are *not* compatible, since in general $\langle Bel \rangle_A$ and $\langle Bel_{\oplus A} \rangle$ are not equal.

This incompatibility of Dempster's rule with the interpretation of belief functions as lower probability functions will come up again in the discussion on the justification of DS theory in the next section.

Exercise 4.3.1 Let $\Omega = \{a, \neg a\}$. The mass functions m_1 and m_2 over Ω are given by $m_1(\{a\}) = 0.7$, $m_1(\Omega) = 0.3$, $m_2(\{a\}) = 0.8$ and $m_2(\Omega) = 0.2$. Describe the mass function $m_1 \oplus m_2$ and the belief function $Bel_1 \oplus Bel_2$.

Exercise 4.3.2 Let $\Omega = \{b, \neg b\}$. The mass functions m_1 and m_2 over Ω are given by $m_1(\{b\}) = 0.7$, $m_1(\Omega) = 0.3$, $m_2(\{\neg b\}) = 0.8$ and $m_2(\Omega) = 0.2$. Describe the mass function $m_1 \oplus m_2$ and the belief function $Bel_1 \oplus Bel_2$.

Exercise 4.3.3 Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. The mass functions m_1 and m_2 over Ω are given as follows. $m_1(\{1, 3, 5\}) = 0.4$, $m_1(\{1, 3\}) = 0.3$, $m_1(\Omega) = 0.3$, $m_2(\{1, 2, 3\}) = 0.6$ and $m_2(\{4, 5, 6\}) = 0.4$. Describe the mass function $m_1 \oplus m_2$ and the belief function $Bel_1 \oplus Bel_2$.

Exercise 4.3.4 Prove proposition 4.3.1.

Exercise 4.3.5 Let Bel_1 and Bel_2 be belief functions over Ω. Show that if $Bel_1 \oplus Bel_2$ is defined and Bel_1 is Bayesian, then $Bel_1 \oplus Bel_2$ is also Bayesian.

Exercise 4.3.6 Let Bel be a belief functions over Ω. Show that $Bel \oplus 1_\Omega$ is defined and that $Bel \oplus 1_\Omega = Bel$.

Exercise 4.3.7 Prove proposition 4.3.4.

4.4 On the Justification of DS Theory

In the previous section, a number of properties of Dempster's combination rule are mentioned. An important property which is in general *not* valid is *idempotency* ($Bel \oplus Bel = Bel$). It follows that Dempster's rule is not universally applicable.

Example 4.4.1 Consider the belief function Bel_1 representing the first piece of evidence of example 4.2.1. Combining this piece of evidence with itself should not change one's degrees of belief in the proposition that the safecracker is left-handed. However, $Bel_1 \oplus Bel_1(L) = 0.91 \neq 0.7 = Bel_1(L)$.

One can also argue against the universal validity of Dempster's rule by pointing out that the rule combines any pair of probability functions, i.e., Bayesian belief functions, over Ω into a combined probability function, and Neapolitan's result of chapter 1 (proposition 1.4.1) shows that such a combination function does not exists.

In his *A Mathematical Theory of Evidence* Shafer mentions two requirements which have to be satisfied when Dempster's rule is applied, namely "that the belief functions to be combined are actually based on entirely distinct bodies of evidence and that the frame of discernment discerns the relevant interaction of these bodies of evidence". (*Glenn Shafer, A Mathematical Theory of Evidence, p. 57.*)

We first briefly discuss the second requirement, namely that the frame has to discern the relevant interaction of the bodies of evidence. Notice that the frame over which the evidence is evaluated is not determined in advance. Agents which are interested in different questions, may choose different frames.

Example 4.4.2 Consider example 4.2.1 of the Safecracker. One can imagine that some insurance company is not as much interested in identifying the individual responsible for cracking the safe as in finding out whether it was an inside job or not. The people of this company might evaluate the evidence over the frame $\{I, \overline{I}\}$ instead of the frame $\Omega = \{x : x \text{ is a possible suspect}\}$ used in example 4.2.1. However, it turns out that Dempster's rule does not always admit such a coarsening of the frame.

Shafer's requirement implies that the choice of a frame should not only take into account the question one is interested in, but also the available pieces of evidence and their possible interactions. The following example illustrates the problems that can arise if the requirement is not satisfied.

Example 4.4.3 Consider the example 4.2.1 of the Safecracker using the frame $\Theta = \{L \cap I, \overline{L \cap I}\}$. Then both Bel_1 and Bel_2 are vacuous, and so is $Bel_1 \oplus Bel_2$, in spite of the support which the two clues taken together give to $L \cap I$. The frame Θ cannot discern this support, because it can only discern the support for L of the first clue and the support for I of the second clue as support for Θ, and $\Theta \cap \Theta = \Theta \neq L \cap I$.

Shafer has given an exact formulation of his requirement that frames have to be sufficiently fine to discern the relevant interaction of the bodies of evidence. We will not repeat this exact formulation. Instead, the reader may assume from now on that frames satisfy this requirement.

The requirement that the belief functions to be combined have to be based on entirely distinct bodies of evidence is formulated rather vaguely and the meaning of this requirement has in fact never been made completely clear. Below we will give an adequate formalisation of this requirement in the context of Dempster's model of belief functions.

The requirement obviously prohibits the use of Dempster's rule to combine a piece of evidence E with itself, or more generally, with some body of evidence E' which is implied by E. (If E implies E', then the evidence E' is already taken into account by the mass function representing E.)

A natural following case to be considered is that where, although E does not (logically) imply E', E considerably increases the probability of E', i.e., $P(E'|E) \gg P(E')$. It might seem plausible that also in this case the effect of the body of evidence E' is (for a large part, at least) already taken into account by the mass function representing E. This may suggest that the rule should only be applied if the bodies of evidence are (probabilistically) independent in the sense that $P(E \cap E') = P(E) \cdot P(E')$.

Indeed, Shafer's requirement that the belief functions have to be based on entirely distinct bodies of evidence pieces of evidence is often interpreted in the literature as meaning that the pieces of evidence have to be probabilistically independent. However, this interpretation is incorrect. In fact, Neapolitan's result of chapter 1 would still disqualify Dempster's rule if only probabilistic independence of the evidence would be required.

In DS theory, there is an other, more relevant concept of independence, which we call *DS-independence*. Roughly speaking, bodies of evidence are called DS-independent if the behaviour of their sources is probabilistically independent. In the context of Dempster's model, an exact definition of the notion of DS-independence can be given. First we give an example to illustrate the difference between the two notions of independence.

Example 4.4.4 Consider some valve V in a nuclear power station. V has two possible states, namely "open" and "closed". The information about the state of the valve V available in the control room derives from two sensors S_1 and S_2. The sensor S_1 functions reliably 90% of the time, and the sensor S_2 functions reliably 80% of the time. If a sensor does not function reliably, then it indicates either that the valve is open or that the valve is closed, but this indication is in no way causally related to the actual state of the valve.

Let r_i : S_i is reliable, and let E_i : S_i indicates that V is open. Suppose both sensors indicate that V is open. The available evidence E_1 and E_2 is probabilistically independent iff $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$, whereas the evidence is DS-independent iff r_1 and r_2 are probabilistically independent.

Let us formalise the evidence of the first sensor of the above example in terms of Dempster's model. The frame corresponding to the target information level is $\Omega = \{o, \neg o\}$, where o : valve V is open. The information on the original information level concerning the sensor S_1 is represented by the sample space $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$, where $\Theta_1 = \{r_1, \neg r_1\}$ and P_1 is given by $P_1(\{r_1\}) = 0.9$. The evidence E_1 is represented by the multivalued mapping $\Gamma_1 : \Theta_1 \to 2^\Omega \backslash \{\emptyset\}$ given by $\Gamma_1(r_1) = \{o\}$ and $\Gamma_1(\neg r_1) = \Omega$.

Since the effect of the evidence E_1 on Ω can only be calculated if the sample space $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$ is known, one might say that the evidence is represented by the *pair* consisting of the multivalued mapping and the original information level. However, we prefer to say that the evidence is represented by Γ_1 and that $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$ is the underlying sample space.

To give a formal definition of DS-independence we need the notion of product probability function.

Definition 4.4.1 *Let* $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$ *and* $\langle \Theta_2, 2^{\Theta_2}, P_2 \rangle$ *be probability spaces. Their* product probability space *is the probability space* $\langle \Theta_1 \times \Theta_2, 2^{\Theta_1 \times \Theta_2}, P \rangle$, *where* P *is given as follows.*

For every $\langle \theta_1, \theta_2 \rangle \in \Theta_1 \times \Theta_2$, $P(\{\langle \theta_1, \theta_2 \rangle\}) = P_1(\{\theta_1\}) \cdot P_2(\{\theta_2\})$.

The thus defined probability function P *is called the* product probability function of P_1 and P_2.

If the probability spaces $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$ and $\langle \Theta_2, 2^{\Theta_2}, P_2 \rangle$ represent independent experiments, then their product probability space is the legitimate representation of the combined experiment.

Definition 4.4.2 *Let Ω be a sample space. Assume that the evidence E_1 is represented by the multivalued mapping $\Gamma_1 : \Theta_1 \to 2^\Omega \backslash \{\emptyset\}$ with underlying probability space $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$, and that the evidence E_2 is represented by $\Gamma_2 : \Theta_2 \to 2^\Omega \backslash \{\emptyset\}$ with underlying probability space $\langle \Theta_2, 2^{\Theta_2}, P_2 \rangle$. The pieces of evidence E_1 and E_2 are called* DS-independent *iff the probability function over $\Theta_1 \times \Theta_2$ is the product probability function of P_1 and P_2.*

Thus the pieces of evidence E_1 and E_2 are DS-independent, not if E_1 and E_2 are independent, but if the underlying probability spaces, representing the *sources* of the evidence, are independent. The property of DS-independence does *not* rely on the multivalued mappings.

The above definition of DS-independence in the context of Dempster's model is based on Shafer's explanation of what he meant by "entirely distinct bodies of evidence". This explanation presupposes an interpretation of DS theory which is essentially Dempster's model. It still remains unclear how Shafer's requirement can be explained without reference to the original information level.

Having clarified the requirements in the context of Dempster's model, it makes sense to ask whether the requirements are sufficient to guarantee the validity of Dempster's combination rule. The answer is negative, since some (rather unrealistic) additional assumptions are needed. Below we give a (slightly simplified) description of these additional assumptions.

Definition 4.4.3 *Let $P \in PROB(\Omega)$ and $E \subseteq \Omega$ such that $P(E) > 0$. The subsets $A_1, A_2, \ldots, A_n$ of Ω are called* equally confirmed by E *iff there exists a λ such that for all $i \in \{1, 2, \ldots, n\}$, $P(A_i|E) = \lambda \cdot P(A_i)$.*

Assume that the evidence E_1 is represented by the multivalued mapping $\Gamma_1 : \Theta_1 \to 2^\Omega$ with underlying probability space $\langle \Theta_1, 2^{\Theta_1}, P_1 \rangle$, and that the evidence E_2 is represented by $\Gamma_2 : \Theta_2 \to 2^\Omega$ with underlying probability space $\langle \Theta_2, 2^{\Theta_2}, P_2 \rangle$.

A pair $\langle \theta_1, \theta_2 \rangle \in \Theta_1 \times \Theta_2$ is called *compatible* iff $\Gamma_1(\theta_1) \cap \Gamma_2(\theta_2) \neq \emptyset$. Dempster's rule needs the additional assumption that every compatible pair $\langle \theta_1, \theta_2 \rangle \in \Theta_1 \times \Theta_2$ is equally confirmed by the evidence $E_1 \cap E_2$.

Example 4.4.5 Consider example 4.4.4. In this case, the additional assumption needed for Dempster's rule implies that the evidence E_1 and E_2 equally confirms the pairs $\langle r_1, r_2 \rangle$, $\langle r_1, \neg r_2 \rangle$, $\langle \neg r_1, r_2 \rangle$, and $\langle \neg r_1, \neg r_2 \rangle$.

This assumption is not very realistic, since the fact that the two sensors agree tends to confirm the hypothesis that the sensors are reliable more than the hypothesis that the sensors are unreliable.

We may conclude that under the interpretation of a belief function as a special kind of lower probability function, Dempster's combination rule is not valid, not even when the bodies of evidence are DS-independent and the frame discerns the relevant interactions of the evidence.

This leaves open the question whether DS theory, and Dempster's rule in particular, can be justified under an interpretation of the formalism which is not based on (generalised) probability theory.

Shafer's original interpretation in *A Mathematical Theory of Evidence* is supported by a reasonable intuitive explanation of the notions involved. Moreover, some examples are given in which the theory gives intuitively appealing results. However, it is telling that Shafer had to appeal to Dempster's model in order to explain his requirement that Dempster's rule should only be applied to "entirely distinct bodies of evidence".

Presently, the *Transferable Belief Model* (TBM) of Philippe Smets is the best developed interpretation of (the formalism of) DS theory in the line of Shafer's original interpretation. In the TBM there exists no presupposed link between belief functions and probability theory. The impact of a piece of evidence is represented by a mass function and Dempster's rule of conditioning is used for the transfer of belief.

Smets has given an axiomatic justification of TBM, that is, a number of principles, which are defended as plausible principles, and which are shown to be necessary and sufficient for degrees of belief to be represented by belief functions and for conditioning to be represented by Dempster's rule of conditioning.

A detailed treatment of Smets' result is beyond the scope of these notes. Instead, we end our discussion with an example (deriving from Smets) which has been used (by different authors) to argue both *for* and *against* Dempster's rule.

Example 4.4.6 Mr. Jones has been murdered by one of the assassins Albert, Bob, and Cindy under orders of Big Boss, who has chosen between these three possible killers as follows. He decided between a male or a female killer by means of tossing a fair coin. A male killer was chosen in case the coin landed heads. Otherwise, a female killer was chosen. No information is available on how he decided between the two male assassins in case the coin landed heads.

Based on the information above, it seems reasonable to say that the killer being male and the killer being female are equally likely. Now suppose that you learn that at the time of the murder, Albert was at the police station, where he was questioned about some other crime. So you can rule out Albert as the killer. Are the killer being male and the killer being female still equally likely, or not? The answer of DS theory (and the TBM) is "yes", whereas generalised probability theory answers "no".

Let $\Omega = \{a, b, c\}$, where a : Albert is the killer, et cetera. The information about the selection process of Big Boss is represented in DS theory by the mass function m given by $m(\{a, b\}) = m(\{c\}) = 0.5$, and in generalised probability theory by the set $\Pi = \{P \in PROB(\Omega) : P(\{a, b\}) = P(\{c\}) = 0.5\}$. So far, there is no disagreement, since $Bel = \Pi_{low}$.

Now consider the effect of the evidence that Albert is not the killer, that is, condition on $\{b, c\}$. In DS theory, this results in the mass function $m_{\oplus\{b,c\}}$ given by $m_{\oplus\{b,c\}}(\{b\}) = m_{\oplus\{b,c\}}(\{c\}) = 0.5$. Thus the killer being male and the killer being female are still considered to be equally likely.

On the other hand, $\Pi_{\{b,c\}} = \{P_{\{b,c\}} : P \in \Pi, P(\{b, c\}) \neq 0\}$, and it easy to check that $(\Pi_{\{b,c\}})_{low}(\{b\}) = 0$, $(\Pi_{\{b,c\}})_{up}(\{b\}) = 0.5$, $(\Pi_{\{b,c\}})_{low}(\{c\}) = 0.5$, and $(\Pi_{\{b,c\}})_{up}(\{c\}) = 1$. Thus generalised probability theory leaves open the possibility that the killer being female is more likely than the killer being male.

Some authors (including Smets) maintain that in the example above the answer of DS theory is to be preferred. Other researchers (including Fagin, Halpern, and the author of these notes) prefer the answer given by generalised probability theory.

Exercise 4.4.1 Let Bel be a belief function over Ω with n focal elements. Show that $Bel \oplus Bel = Bel$ iff the focal elements of Bel are pairwise disjoint and $m(A) = \frac{1}{n}$, for each focal element A.

Exercise 4.4.2 Specify the product probability space of the probability spaces underlying the pieces of evidence of example 4.4.4. How should in your opinion the probabilities be updated after combining the evidence? How are these probabilities (implicitly) updated when using Dempster's rule of combination?

Exercise 4.4.3 Let $P \in PROB(\Omega)$ and let $A_1, A_2, \ldots, A_n$ and E be subsets of Ω with positive probability. Show that $A_1, A_2, \ldots, A_n$ are equally confirmed by E iff for all $i, j \in \{1, 2, \ldots, n\}$, $P(E|A_i) = P(E|A_j)$.

4.5 Applications of DS Theory

In spite of the problems with justifying DS Theory, it is still occasionally applied in several different AI domains. (Medical diagnosis, computer vision, locating submarines, classifying radar images, robot localisation, etc.) People using DS Theory mention the following reasons:

1. The difficult problem of specifying all (prior) probabilities can be avoided.
2. In addition to uncertainty, also ignorance can be expressed.
3. It is straightforward to express pieces of evidence with different levels of abstraction.
4. Dempster's combination rule can be used to combine pieces of evidence.

The first three reasons are related and can be considered advantages of DS Theory relative to Probability Theory, but the same advantages can be achieved by Generalised Probability Theory.

Generalised Probability Theory lacks a (simple) combination rule, but such a rule is hard to justify anyway. Moreover, in concrete applications, often not much serious attention is paid to the assumptions required for applying Dempster's rule. But it is only fair to mention that also in applications of (Generalised) Probability Theory often many simplifying assumptions are made (explicitly or implicitly) without proper justification.

As mentioned in the first section, the availability of a combination rule has computational advantages, but DS Theory still has potential computational complexity problems, since its mass functions are potentially exponentially larger than probability functions. Another disadvantage of DS Theory is that it still lacks a well-established decision theory, whereas Bayesian Decision Theory (maximising expected utility) is almost universally accepted by those using Bayesian Probability Theory. Generalised Probability Theory has the same problems, or worse.

Experimental comparisons between DS Theory and Probability Theory are scarcely performed and it is usually difficult to draw hard conclusions from the experimental results, since typically several assumptions have to be made when formalising the problem. Therefore, if a particular theory gives better experimental results, this does not necessarily mean that the theory is superior, since it can also be caused by a suboptimal analysis of the problem in terms of the other formalisms.