

Uncertainty Analysis using Belief Functions:  
Applications to Statistical Inference and  
Pattern Recognition

Thierry Denceux

May 3, 2021



# Contents

<b>1</b>	<b>Classical models</b>	<b>11</b>
1.1	Uncertainty . . . . .	12
1.2	Set-based representation of uncertainty . . . . .	13
1.2.1	Operations on sets . . . . .	13
1.2.2	Relationship with propositional logic . . . . .	17
1.2.3	Necessity and possibility functions . . . . .	18
1.2.4	Limitations of sets for representing partial information . . . . .	19
1.3	Probability . . . . .	20
1.3.1	Basic definitions . . . . .	20
1.3.2	Interpretations of probability . . . . .	22
1.3.3	Operations on probabilities . . . . .	24
1.3.4	Justification of probabilities: Cox's theorem . . . . .	26
1.3.5	Principle of indifference . . . . .	27
1.4	Conclusions . . . . .	29
<b>I</b>	<b>Fundamentals</b>	<b>31</b>
<b>2</b>	<b>Representation of Evidence</b>	<b>33</b>
2.1	Mass function . . . . .	33
2.1.1	Definitions . . . . .	33
2.1.2	Semantics . . . . .	34
2.2	Belief and plausibility functions . . . . .	37
2.2.1	Definitions . . . . .	37
2.2.2	Properties . . . . .	38
2.2.3	Vector representation . . . . .	40
2.3	Special cases and related theories . . . . .	41
2.3.1	Bayesian mass functions . . . . .	41
2.3.2	Consonant mass functions . . . . .	42

2.3.3	Relation with imprecise probabilities . . . . .	44
<b>3</b>	<b>Combination of independent items of evidence</b>	<b>47</b>
3.1	Introductory example . . . . .	47
3.2	Dempster's rule . . . . .	49
3.2.1	Definition and elementary properties . . . . .	49
3.2.2	Commonality function . . . . .	51
3.2.3	Conditioning . . . . .	54
3.2.4	Computational complexity . . . . .	56
3.3	Related combination rules . . . . .	57
3.4	Separable belief functions . . . . .	60
<b>4</b>	<b>Uncertainty Principles</b>	<b>65</b>
4.1	Inclusion relations . . . . .	66
4.1.1	Belief and commonality-based inclusion relations . . . . .	66
4.1.2	Strong inclusion . . . . .	68
4.1.3	Weight-based inclusion . . . . .	70
4.2	Uncertainty measures . . . . .	71
4.2.1	Nonspecificity . . . . .	71
4.2.2	Conflict measures . . . . .	73
4.2.3	Total uncertainty measures . . . . .	74
4.3	Applications . . . . .	75
4.3.1	Least committed belief function from consonant sets . . . . .	76
4.3.2	Conditional embedding . . . . .	76
4.3.3	Partial belief specifications . . . . .	77
4.3.4	Combination of mass functions with unknown dependence . . . . .	77
<b>5</b>	<b>The cautious rule of combination</b>	<b>81</b>
<b>6</b>	<b>Varying the granularity of the frame</b>	<b>83</b>
6.1	Refinement and coarsening . . . . .	83
6.2	Special case of product spaces . . . . .	85
6.2.1	Marginalization and vacuous extension . . . . .	85
6.2.2	Application to evidential reasoning . . . . .	86
6.3	Conditioning and deconditioning . . . . .	87
6.4	Applications . . . . .	88
6.4.1	Discounting . . . . .	88
6.4.2	Generalized Bayes' Theorem . . . . .	89

<b>7</b>	<b>Decision-making</b>	<b>93</b>
7.1	Formal framework . . . . .	93
7.2	Elements of classical decision theory . . . . .	95
7.2.1	Decision-making under complete ignorance . . . . .	95
7.2.2	Decision-making with probabilities . . . . .	99
7.2.3	Savage's theorem . . . . .	101
7.2.4	Ellsberg's paradox . . . . .	103
7.3	Decision-making with belief functions . . . . .	104
7.3.1	Upper and lower expected utility . . . . .	106
7.3.2	Other approaches . . . . .	108
7.3.3	Axiomatic justifications . . . . .	111
<b>8</b>	<b>Belief Functions on Infinite Spaces</b>	<b>119</b>
8.1	General definitions and results . . . . .	119
8.1.1	Definitions . . . . .	120
8.1.2	Belief function induced by a source . . . . .	120
8.1.3	Dempster's rule . . . . .	123
8.2	Practical models . . . . .	124
8.2.1	Consonant random closed sets . . . . .	124
8.2.2	Random closed intervals . . . . .	125
<b>II Application to Machine Learning and Statistical Inference</b>		<b>129</b>
<b>9</b>	<b>Statistical Inference</b>	<b>131</b>
9.1	Limitations of classical approaches . . . . .	132
9.1.1	Frequentist approach . . . . .	132
9.1.2	Bayesian approach . . . . .	134
9.1.3	Likelihood-based approach . . . . .	135
9.2	Dempster's method . . . . .	137
9.2.1	General method . . . . .	137
9.2.2	Application to a Bernoulli sample . . . . .	139
9.3	Likelihood-based method . . . . .	141
9.3.1	General method . . . . .	141
9.3.2	Bernoulli example . . . . .	143
9.3.3	Properties . . . . .	144
9.4	Prediction . . . . .	145
9.4.1	General method . . . . .	146
9.4.2	Bernoulli example . . . . .	147

9.4.3	Monte Carlo approximation . . . . .	149
9.4.4	Relationship with the Bayesian posterior predictive distribution . . . . .	150
<b>10</b>	<b>Uncertain Data</b>	<b>153</b>
<b>11</b>	<b>Classification</b>	<b>155</b>
<b>12</b>	<b>Clustering</b>	<b>157</b>
<b>13</b>	<b>Belief functions on sets of relations</b>	<b>159</b>

# List of Figures

1.1	Refinement of a frame of discernment. . . . .	14
1.2	Cylindrical extension of a subset $A$ of $\Omega_X$ in $\Omega_X \times \Omega_Y$ (a), and projection of a relation $R$ on $\Omega_X$ (b). . . . .	16
1.3	Reasoning with sets. Knowing that $X \in A$ and the relation $R$ constraining the values of $X$ and $Y$ , we can deduce that $Y \in B$ , where $B$ is the projection on $\Omega_X$ of the intersection of $R$ with the cylindrical extension of $A$ . . . . .	17
2.1	Random code setup. . . . .	36
2.2	Belief and plausibility functions (Example 2.3). . . . .	38
3.1	Combination of evidence in the murder example. . . . .	48
6.1	Refinement of a frame of discernment. . . . .	84
6.2	Conditional embedding operation. The mass on $m^X(A B)$ is transferred to $(A \times \Omega_Y) \cup (\Omega_X \times \bar{B})$ . . . . .	88
9.1	Jeffreys prior for a Bernoulli sample. . . . .	136
9.2	Focal sets $(\{0\} \times [0, W]) \cup (\{1\} \times [W, 1])$ of the belief function $Bel^{\mathcal{X} \times \Theta}$ in the case of a Bernoulli sample. . . . .	140
9.3	Contour functions (normalized likelihood functions) for the binomial distribution with $\hat{\theta} = 0.4$ and $n \in \{10, 20, 100\}$ . . . . .	144
9.4	Three cases in the computation of the predictive belief func- tion on $Y$ in the Bernoulli example. . . . .	147
9.5	Predictive belief and plausibility of success for a Bernoulli trial based on the contour function $pl_x(\theta)$ on the probability of success $\theta$ . . . . .	149





# List of Tables

1.1	Dutch book argument; gains in the three bets. . . . .	24
2.1	Four mass functions on $\Omega = \{a, b, c\}$ in Example 2.1. . . . .	34
2.2	Binary ordering in the case $ \Omega  = 3$ . . . . .	40
7.1	Payoff matrix (in €) for the investment example. . . . .	94
7.2	Outcomes in Ellsberg's paradox. . . . .	104
7.3	Lower and upper expected utilities in Ellsberg's urn experiment. . . . .	108



## Chapter 1

# Preliminaries: classical models of partial information

This book is about the *theory of belief functions*, a formal framework for reasoning and making decisions under uncertainty. This framework originates from Arthur Dempster's seminal work on statistical inference with lower and upper probabilities [18, 20]. It was then further developed by Glenn Shafer [77] who showed that belief functions can be used as a general framework for representing and reasoning with uncertain information, beyond the very important but limited confines of statistical inference. The theory of belief functions, also referred to as *Evidence theory* or *Dempster-Shafer theory*, has been widely used in several fields such as artificial intelligence, information fusion, pattern recognition, etc. Recently, there has been a revived interest in its application to statistical inference. This formalism seems particularly well suited to situations where we are facing limited information such as low quality data or partially reliable expert opinions. It has been applied to thousands of real-world problems in a wide range of application areas, including robotic perception, medicine, target detection and tracking, risk analysis, auditing, economics and finance, etc.

In this introductory chapter, we will discuss the concept of uncertainty and review two popular formalisms for handling uncertainty: sets and probabilities. As we shall see all along this book, the theory of belief functions builds upon these two approaches: in a way, a belief function can be seen as the assignment of probabilities to sets.

## 1.1 Uncertainty

The concept of uncertainty has been well described by Dennis Lindley in his book “Understanding uncertainty” [57]:

*“There are some things that you know to be true, and others that you know to be false; yet, despite this extensive knowledge that you have, there remain many things whose truth or falsity is not known to you. We say that you are uncertain about them. You are uncertain, to varying degrees, about everything in the future; much of the past is hidden from you; and there is a lot of the present about which you do not have full information. Uncertainty is everywhere and you cannot escape from it.”*

This description highlights several important aspects of uncertainty. First, it is a consequence of lack of knowledge. You are uncertain about a proposition if you do not know whether it is true or false. Equivalently, we can say that you are uncertain about some question if you do not know its correct answer. Secondly, uncertainty is a matter of degree. Typically, you may have *some* information about a question, but this information may not be sufficient to identify the correct answer. Consider, for instance, the temperature in the room where you are sitting while you are reading these lines. If you do not have a precision thermometer next to you, you are uncertain about this temperature. However, your senses provide you with some information about it (you may know for sure that it is somewhere in the range 15-40 degrees Celsius). How to represent such partial information and use it to draw conclusions and make decisions is the topic of this book. As also emphasized by Lindley, uncertainty is ubiquitous – one could even hold the view that uncertainty is the rule, while complete certainty is the exception. Modeling uncertainty is thus a topic of considerable interest.

As a first step towards a formalization of uncertainty, one needs to make some assumptions and define some notations. We suppose that we are concerned with some question  $Q$ , for which there exists one and only one correct answer  $X$ , which is assumed to be an element of some set  $\Omega$ . Following Shafer [77], we will call  $\Omega$  the *frame of discernment*. We can remark that we have already made two important assumptions here: there is only one correct answer, and it is surely contained in some known set  $\Omega$ . The second assumption, in particular, may be questionable for some problems. For instance, in risk analysis, we typically enumerate possible scenarios. How can we be sure that no scenario has been forgotten? In effect, many industrial disasters occur because some scenario has not been foreseen. There are, however, many cases where the frame of discernment is exhaustive by construction. For instance, if  $Q$  concerns a temperature expressed in Kelvin, it is surely

contained in the real interval  $[0, +\infty)$ .

For the moment, we will leave aside potential difficulties related to the definition of the frame of discernment, and we will assume that it can be defined unambiguously for the question  $Q$  of interest. For any subset  $A$  of  $\Omega$ , the proposition “the correct answer to  $Q$  is in  $A$ ” will be denoted as “ $X \in A$ ”. For instance, if  $X$  is the room temperature, the proposition “ $X$  is in the range 15-40 degrees Celsius” will be denoted as “ $X \in [15, 40]$ ”. Sometimes, a proposition “ $X \in A$ ” will be unambiguously identified to the subset  $A$  itself, and we will loosely speak of “proposition  $A$ ”.

In the rest of this introductory chapter, we will briefly review two classical models of partial information: sets, and probabilities. A good understanding of these two models is important, as belief functions – the main topic of this book – combine and subsume these two models. Most of the set-theoretic or probabilistic concepts outlined here will have extensions in the theory of belief functions.

## 1.2 Set-based representation of uncertainty

Perhaps the simplest way of representing partial information about some question is as a set  $A \subseteq \Omega$  that certainly contains the true answer  $X$ . There is a vast literature on set-membership approaches to uncertainty, with application, e.g., in computer science and automatic control [63, 64, 46]. An important special case is interval arithmetics, which is based on syntactic rules to compute with intervals, making it possible to produce rigorous enclosures of solutions to model equations. The fundamental operations on sets are simple forms of operations on the more complex representations that will be studied, in particular, in Chapters 3 and 6. We will briefly review these operations in the following section, highlight the link with propositional logic, and introduce the important concepts of possibility and necessity measures. Finally, we will discuss some limitations of this approach.

### 1.2.1 Operations on sets

As we will see later in this course, a major task when reasoning with uncertainty is *information fusion*, i.e., combining pieces of information (evidence) from different sources. Assume that two sources provide two subsets  $A$  and  $B$  of  $\Omega$ , assumed to contain the answer to the question of interest. How to combine these pieces of information?

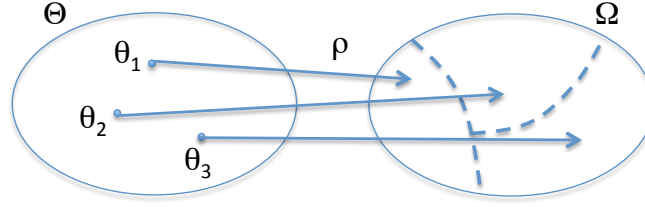


Figure 1.1: Refinement of a frame of discernment.

**Intersection, union.** If both sources can be trusted, then it is reasonable to consider that the true answer is in the intersection of  $A$  and  $B$ , denoted by  $A \cap B$ , which is the set containing the elements of  $\Omega$  that belong to both  $A$  and  $B$ . This mode of fusing information is called *conjunctive*; it is reasonable when all information sources are assumed to be reliable. However, when  $A$  and  $B$  are disjoint, i.e.,  $A \cap B = \emptyset$ , this rule leads a contradiction. In that case, the assumption that the two sources can be trusted is no longer tenable. It is then more cautious to conclude that the true answer is in the union of  $A$  and  $B$ , denoted by  $A \cup B$ , which is the set containing the elements of  $\Omega$  that belong to  $A$  or to  $B$ . This is the simplest form of *disjunctive* rule for pooling information, which is suitable when at least one of the sources is assumed to be reliable.

**Inner and outer approximation** We have already noticed that the construction of the frame of discernment is based on the knowledge we have at a given time concerning possible answers to some question. At some point of the modeling process, it may be convenient to refine the frame, i.e., to break down some possibilities into more detailed ones. Formally, the notion of refinement of a frame can be defined as follows [77].

Let  $\Omega$  and  $\Theta$  be two frames of discernment. We recall that a *partition* of a set  $\Theta$  is a collection of pairwise disjoint subsets of  $\Theta$ , the union of which is equal to  $\Theta$ . We say that  $\Omega$  is a *refinement* of  $\Theta$  (or, equivalently,  $\Theta$  is a *coarsening* of  $\Omega$ ) if there is a one-to-one mapping  $\rho$  between the singletons of  $\Theta$  and a partition of  $\Omega$  (see Figure 6.1). This mapping can be extended to the whole power set of  $\Theta$  as follows,

$$\forall A \subseteq \Theta, \quad \rho(A) = \bigcup_{\theta \in A} \rho(\{\theta\}). \quad (1.1)$$

Assume that some information about  $X$  is modeled by a subset  $A$  of  $\Theta$ . The same information can be expressed without loss of information in the

finer frame  $\Omega$  as  $\rho(A)$ . We say that no information has been lost, because we can always retrieve  $A$  from  $B = \rho(A)$  as follows,

$$A = \{\theta \in A \mid \rho(\{\theta\}) \subseteq B\}. \quad (1.2)$$

Consider now the other situation, in which we have a subset  $B$  of  $\Omega$  representing some information about  $X$ . Usually, there is no subset  $A$  of  $\Theta$  such that  $\rho(A) = B$ , so that the information expressed in  $\Omega$  does not have an exact representation in  $\Theta$ . However, we can define approximate representations as follows. Let  $\underline{\rho}^{-1}$  and  $\bar{\rho}^{-1}$  be two mappings from  $2^\Omega$  to  $2^\Theta$  defined as

$$\underline{\rho}^{-1}(B) = \{\theta \in \Theta \mid \rho(\{\theta\}) \subseteq B\}, \quad (1.3a)$$

$$\bar{\rho}^{-1}(B) = \{\theta \in \Theta \mid \rho(\{\theta\}) \cap B \neq \emptyset\}, \quad (1.3b)$$

for any subset  $B$  of  $\Omega$ . The subsets  $\underline{\rho}^{-1}(B)$  and  $\bar{\rho}^{-1}(B)$  are called, respectively, the *inner reduction* and the *outer reduction* of  $B$  [77]. When computing the image by  $\rho$  of the inner or the outer reduction of  $B$ , we do not recover  $B$  in general. Let  $\underline{B} = \rho[\underline{\rho}^{-1}(B)]$  and  $\bar{B} = \rho[\bar{\rho}^{-1}(B)]$ . The following relations hold:

$$\underline{B} \subseteq B \subseteq \bar{B}, \quad (1.4)$$

and the inclusions may be strict. The sets  $\underline{B}$  and  $\bar{B}$  are called, respectively, the inner and outer approximation of  $B$  induced by the partition of  $\Omega$ , and the pair  $(\underline{B}, \bar{B})$  is called a *rough set* [68, 69]. Rough sets have been extensively studied in the context of data mining, and some connections to belief functions have been explored (see, e.g., [115]).

**Projection, extension.** Let us now assume that we have two questions of interest, whose true answers are denoted by  $X$  and  $Y$  ( $X$  and  $Y$  may be called *variables*). Let  $\Omega_X$  and  $\Omega_Y$  be the frames of discernment for  $X$  and  $Y$ . To represent information about the values that  $X$  and  $Y$  may take jointly, we need to place ourselves in the Cartesian product  $\Omega_X \times \Omega_Y$ , denoted more concisely by  $\Omega_{XY}$ , and defined as the set of ordered pairs  $(x, y)$  of an element of  $\Omega_X$  and an element of  $\Omega_Y$ . A subset of  $R$  of  $\Omega_{XY}$  is called a *relation*. It can be used to represent a constraint on the values that  $X$  and  $Y$  may take jointly.

**Example 1.1** Let  $\Omega_X = \{x_1, x_2, x_3\}$  be a set of symptoms and  $\Omega_Y = \{y_1, y_2, y_3\}$  a set of diseases. The relation  $R = \{(x_1, y_1), (x_1, y_2), (x_3, y_2), (x_2, y_3)\}$  may express that symptom  $x_1$  is associated to diseases  $y_1$  and  $y_2$ ,  $x_3$  is associated to  $y_2$ , and  $x_2$  is associated to  $y_3$ .

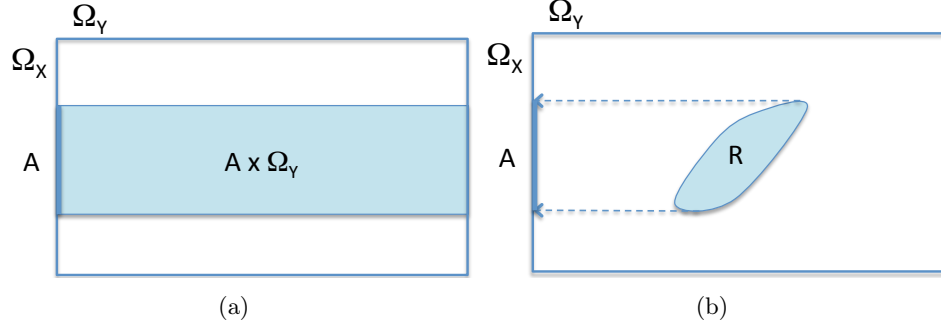


Figure 1.2: Cylindrical extension of a subset  $A$  of  $\Omega_X$  in  $\Omega_X \times \Omega_Y$  (a), and projection of a relation  $R$  on  $\Omega_X$  (b).

It is clear that  $\Omega_{XY}$  is a refinement of both  $\Omega_X$  and  $\Omega_Y$ . For instance, we can define the following refining  $\rho$  from  $2^{\Omega_X}$  to  $2^{\Omega_{XY}}$ :

$$\rho(A) = A \times \Omega_Y, \quad (1.5)$$

for all  $A \subseteq \Omega_X$ . The set  $\rho(A)$  is called the *cylindrical extension* of  $A$  in  $\Omega_{XY}$  and is denoted by  $A \uparrow \Omega_{XY}$  (see Figure 1.2(a)). Conversely, let  $R$  be a subset of  $\Omega_{XY}$ . Its outer reduction is

$$\bar{\rho}^{-1}(R) = \{x \in \Omega_X \mid \rho(\{x\}) \cap R \neq \emptyset\} \quad (1.6a)$$

$$= \{x \in \Omega_X \mid \exists y \in \Omega_Y, (x, y) \in R\}. \quad (1.6b)$$

This set is denoted by  $B \downarrow \Omega_X$  and is called the *projection* of  $R$  on  $\Omega_X$  (see Figure 1.2(b)).

To see how these notions can be used in a reasoning process, assume that we have

- Evidence that  $X$  belongs to a subset  $A$  of  $\Omega_X$ ;
- Evidence about the values that  $X$  and  $Y$  can take jointly, represented by a relation  $R \subseteq \Omega_{XY}$  (see Figure 1.3).

What can we deduce about  $Y$ ? Let  $B$  denote the set of possible values for  $Y$ . It is clear that  $y$  belongs to  $B$  if and only if there is some  $x$  in  $A$  such that  $(x, y) \in R$ . Formally:

$$B = \{y \in \Omega_Y \mid \exists x \in A, (x, y) \in R\}, \quad (1.7)$$



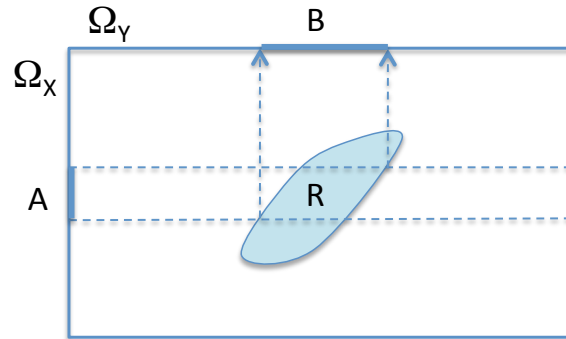


Figure 1.3: Reasoning with sets. Knowing that  $X \in A$  and the relation  $R$  constraining the values of  $X$  and  $Y$ , we can deduce that  $Y \in B$ , where  $B$  is the projection on  $\Omega_X$  of the intersection of  $R$  with the cylindrical extension of  $A$ .

which can be written as

$$B = (R \cap (A \uparrow \Omega_{XY})) \downarrow \Omega_Y. \quad (1.8)$$

This kind of reasoning may straightforwardly be applied to any number of variables. As we will see in Chapter 6, it can be extended to the more complex framework of belief functions.

### 1.2.2 Relationship with propositional logic

Propositional logic is another formalism closely related to set theory. The basic constructs of that formalism are *propositional variables*  $p, q, r, \dots$ , which represent statements (elementary propositions) that can be true ( $T$ ) or false ( $F$ ), and *connectives*  $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$ , which make it possible to build formulas expressing more complex propositions. The meaning of a connective is described by a truth table. For instance, the following table,

$p$	$q$	$p \rightarrow q$
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$T$
$F$	$F$	$T$

states that  $p \rightarrow q$  is true if and only if  $p$  is false, or  $q$  is true.

An *interpretation* is a mapping from the set of propositional variables to the set  $\{T, F\}$  of truth values. To each formula  $\phi$  corresponds the set  $\mathcal{I}(\phi)$  of interpretations under which it is true. For instance, to  $p \rightarrow q$  corresponds the set  $\{(T, T), (F, T), (F, F)\}$ . If  $\phi$  and  $\psi$  are two formulas, then  $\mathcal{I}(\phi \wedge \psi) = \mathcal{I}(\phi) \cap \mathcal{I}(\psi)$ ,  $\mathcal{I}(\phi \vee \psi) = \mathcal{I}(\phi) \cup \mathcal{I}(\psi)$  and  $\mathcal{I}(\neg\phi) = \overline{\mathcal{I}(\phi)}$ , where  $\overline{\mathcal{I}(\phi)}$  denotes the complement of  $\mathcal{I}(\phi)$  in the set of all interpretations. Interpretations can be seen as representing states of the world, and a proposition can be identified to the set of states of the world under which it is true.

Instead of defining a frame of discernment in the first place, as we did in Section 1.1, we might as well have started by defining a set of elementary propositions, large enough to supply the needs of a given subject matter, and a formula  $\phi$  describing initial knowledge. The set  $\mathcal{I}(\phi)$  of interpretations of  $\phi$  can then be seen as a set of possible states of the world, which is equivalent to our frame of discernment  $\Omega$ . For instance, let  $p, q, r$  be 3 propositional variables, each one corresponding to some answer to question  $Q$ . The formula

$$\phi = (p \vee q \vee r) \wedge \neg(p \wedge q) \wedge \neg(p \wedge r) \wedge \neg(q \wedge r) \quad (1.9)$$

expresses the fact that one and only one of the elementary propositions is true. The corresponding set of interpretations is

$$\mathcal{I}(\phi) = \{(T, F, F), (F, T, F), (F, F, T)\}. \quad (1.10)$$

### 1.2.3 Necessity and possibility functions

Any subset  $A$  of  $\Omega$  can be represented by its *indicator function*, defined by

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise,} \end{cases} \quad (1.11)$$

for all  $\omega \in \Omega$ . If the proposition “ $X \in A$ ” is known to be true, then  $\mathbb{1}_A(\omega) = 1$  means that the proposition “ $X = \omega$ ” is *possibly* true. Conversely,  $\mathbb{1}_A(\omega) = 0$  means that the proposition “ $X = \omega$ ” is *certainly* false. The following equalities are straightforward,

$$\mathbb{1}_{A \cap B}(\omega) = \min(\mathbb{1}_A(\omega), \mathbb{1}_B(\omega)) \quad (1.12a)$$

$$\mathbb{1}_{A \cup B}(\omega) = \max(\mathbb{1}_A(\omega), \mathbb{1}_B(\omega)), \quad (1.12b)$$

for all  $\omega \in \Omega$ . Now,  $2^\Omega$  denote the power set of  $\Omega$ , i.e., the collection of all subsets of  $\Omega$ . Let us consider the following functions  $N_A$  and  $\Pi_A$  from  $2^\Omega$  to

$\{0, 1\}$ ,

$$N_A(B) = \begin{cases} 1 & \text{if } A \subseteq B, \\ 0 & \text{otherwise,} \end{cases} \quad (1.13a)$$

$$\Pi_A(B) = \begin{cases} 1 & \text{if } A \cap B \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \quad (1.13b)$$

Functions  $N_A$  and  $\Pi_A$  are called, respectively, necessity and possibility functions. It is clear that we have  $N_A(B) = 1$  iff the proposition “ $X \in A$ ” is certainly true, and  $\Pi_A(B) = 1$  iff it is possibly true. It is also easy to check that the following equalities hold, for all  $B, C \subseteq \Omega$ ,

$$\Pi_A(B) = \max_{\omega \in B} \mathbb{1}_A(\omega), \quad (1.14a)$$

$$N_A(B) = 1 - \Pi_A(\overline{B}), \quad (1.14b)$$

$$\Pi_A(B \cup C) = \max(\Pi_A(B), \Pi_A(C)), \quad (1.14c)$$

$$N_A(B \cap C) = \min(N_A(B), N_A(C)). \quad (1.14d)$$

**Fuzzy sets, possibility theory.** Zadeh [116] has proposed to extend the notion of indicator function, so that it takes values in the interval  $[0, 1]$ . A mapping  $\mu_F : \Omega \rightarrow [0, 1]$  is called the membership function of a *fuzzy set*  $F$ . A fuzzy set can be viewed a set with “blurred”, unsharp boundaries. Each value  $\mu(\omega)$  is interpreted as a degree of membership of  $\omega$  to the fuzzy set. The intersection and union of fuzzy sets are defined by extending (1.12b) to membership functions.

Now, assume that some information is represented by a fuzzy set  $F$ . We can define an induced possibility function  $\Pi$  by generalizing (1.14a) to

$$\Pi_A(B) = \max_{\omega \in B} \mathbb{1}_A(\omega), \quad (1.15)$$

for all  $B \subseteq \Omega$ . Function  $\Pi_A$  then becomes a possibility measure with values in  $[0, 1]$ . The corresponding necessity function can be defined by (1.14b), and it is easy to see that (1.14c) and (1.14d) still hold. The resulting model, called *Possibility Theory* [117, 30], formalizes the idea that possibility and certainty may admit degrees. As we will see in Chapter 2, Possibility Theory has strong links with the theory of belief functions, which is even more general.

#### 1.2.4 Limitations of sets for representing partial information

Crisp (non fuzzy) sets have several limitations as a formalism for representing partial information.

First of all, they do not allow the expression of doubt. As a consequence, they favor a conservative approach, in which the sets have to be chosen very large to ensure that they contain the true value with full certainty. A lot of information is usually lost in such a representation. For instance, an expert is asked to give an interval that surely contains the mean sea level in 2050, they will give a wide interval, even though they may actually believe that the mean sea level will be contained within narrower bounds. By dropping the requirement to define sharp boundaries, fuzzy sets are a possible answer to this problem. As we will show later, belief functions provide an even more general answer, by allowing one to provide different sets (not necessarily nested) with attached degrees of support.

Another limitation of the set-based representation is that it does not allow us to represent statistical evidence. Assume, for instance, that an urn contains 99% of black balls and 1% of white balls. Let  $X \in \Omega = \{\text{black, white}\}$  be the color of a ball  $t$  be drawn from the urn. There is no way, using sets only, to express the fact that the ball is more likely to be black than white. Here, Possibility Theory does not help, because possibility measures defined by (1.15), being non additive, are not suitable for representing proportions or frequencies. To express such knowledge, we need a different paradigm: probability, which will be discussed in the next section.

## 1.3 Probability

Probability theory is another classical formalism for representing and reasoning with uncertainty. After recalling some basic definitions, we will provide a brief review of interpretations and justifications of this approach.

### 1.3.1 Basic definitions

To avoid technical difficulties, let us first assume the frame  $\Omega$  to be finite. A *probability mass function* on  $\Omega$  is a mapping  $p : \Omega \rightarrow [0, 1]$  such that

$$\sum_{\omega \in \Omega} p(\omega) = 1. \quad (1.16)$$

It can be seen as an allocation of masses to the elements of  $\Omega$ , the total mass being equal to unity. The mass  $P(A)$  received by a subset  $A \subseteq \Omega$  is called the probability of  $A$ ,

$$P(A) = \sum_{\omega \in A} p(\omega), \quad (1.17)$$

and the mapping  $P : 2^\Omega \rightarrow [0, 1]$  is called a *probability measure*. It verifies the following properties,

1.  $P(\Omega) = 1$ ;
2. For all elements  $A$  and  $B$  of  $2^\Omega$  such that  $A \cap B = \emptyset$ ,

$$P(A \cup B) = P(A) + P(B). \quad (1.18)$$

We can define the notion of probability measure in a more setting, without assuming  $\Omega$  to be finite. We first need to consider a non-empty collection of subsets of  $\Omega$  (called *events*), closed under complementation and finite union. Such a collection is called an *algebra* of subsets of  $\Omega$ . Let  $\mathcal{A}$  be such collection. We can remark that  $\Omega$  and  $\emptyset$  necessarily belong to  $\mathcal{A}$ . A *finitely additive probability measure* on  $(\Omega, \mathcal{A})$  is a function  $P$  from  $\mathcal{A}$  to  $[0, 1]$  verifying  $P(\Omega) = 1$  and (1.18), for all elements  $A$  and  $B$  of  $\mathcal{A}$ .

We can easily deduce from (1.18) that, for any elements  $A$  and  $B$  of  $\mathcal{A}$ ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (1.19)$$

More generally, we can prove by induction that, for any  $k \geq 2$  and any collection  $A_1, \dots, A_k$  of elements of  $\mathcal{A}$ ,

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} P\left(\bigcap_{i \in I} A_i\right). \quad (1.20)$$

As we will see in Chapter 2, a weaker form of this property characterizes belief functions.

The notion of finitely additive probability is often extended to allow probabilities to be assigned to the union or intersection of *countable* families or events. For this, we need to consider a non-empty collection  $\mathcal{A}$  of subsets of  $\Omega$  that is closed under complementation and countable union. Such a family is called a  $\sigma$ -*algebra*. A *countably additive probability measure* on  $(\Omega, \mathcal{A})$  is a function  $P$  from  $\mathcal{A}$  to  $[0, 1]$  such that  $P(\Omega) = 1$  and, for all countable collections  $(A_i)$ ,  $i = 1, \dots, \infty$ , of pairwise disjoint elements of  $\mathcal{A}$ ,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (1.21)$$

The triple  $(\Omega, \mathcal{A}, P)$  is called a *probability space*. A useful example of a probability space is  $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ , where  $\mathcal{B}(\mathbb{R}^p)$  is the Borel  $\sigma$ -algebra on  $\mathbb{R}^p$ ,

defined as the smallest  $\sigma$ -algebra containing all closed subsets of  $\mathbb{R}^p$ . A probability measure  $P$  on  $(\Omega, \mathcal{A}, P)$  is said to be *absolutely continuous* if there exists an integrable *density function*  $f : \mathbb{R}^p \rightarrow [0, +\infty)$  such that

$$P(A) = \int_A f(x)dx, \quad (1.22)$$

for all  $A \in \mathcal{B}(\mathbb{R}^p)$ . By comparing (1.22) with (1.17), we can see the density function plays the same role here as the probability mass function in the finite case.

### 1.3.2 Interpretations of probability

The mathematical model briefly described above may be used to represent different aspects of the real world. In particular, it can be used to represent *objective* properties of random experiments, or *subjective* degrees of belief. These two interpretations will be briefly reviewed below.

#### Objective probabilities

Probability theory is clearly suitable to represent aleatory uncertainty, in which case the probability  $P(A)$  for an event  $A \subseteq \Omega$  is interpreted either as a *frequency* (actually, the limit of the frequency with which event  $A$  occurs, if the random experiment is repeated  $n$  times and  $n \rightarrow +\infty$ ), or as a *propensity* [71] (i.e., the tendency of  $A$  to happen across a large number of repetitions of the random experiment). Since frequencies are additive, the additivity axiom (1.21) is well justified. Such probabilities can be considered as objective, because they describe physical properties of the chance setup. For instance, when tossing a coin, the probabilities  $P(\text{Heads}) = P(\text{Tails}) = 1/2$  can be deduced from the symmetry of the coin.

#### Subjective probabilities

The use of probability measures to represent epistemic uncertainty (as advocated by the Bayesian school, see, e.g., [11, 47]) is more problematic, because in this case probabilities can clearly no longer be interpreted as frequencies. In this context, they are usually interpreted as subjective (or personal) *degrees of belief*. However, we need to define more precisely the meaning of this notion and to explain why degrees of belief should be additive. This can be done in, at least, two ways: using a constructivist or a behavioral approach.

**Constructivist approach** In the constructivist approach, we construct a probability measure  $P$  by comparing our evidence (i.e., what we know) about  $\Omega$  to a random experiment with known chances [79]. This allows us to construct a scale of degrees of belief, with canonical examples. For instance, in a coin tossing game, the chance for Heads is  $1/2$ , which is taken as our degree of belief that Heads will come up. If our beliefs about the truth of some proposition  $A$  (e.g., “There is life on Mars”) is comparable to our belief that Heads will come up when tossing a coin, we can say that our personal probability for  $A$  is  $1/2$ .

**Betting rates** Another approach, dominant in the literature on subjective probabilities, attaches to probabilities the semantics of betting rates. The so-called “Dutch book” argument <sup>1</sup>, first put forward by Ramsey [73] and de Finetti [15], then shows that consistent betting behavior, in some sense, should be based on probabilities. The Dutch book argument can be summarized as follows.

Assume that you have to enter a game where there is a player and a banker. The player gives an amount of money  $\$p$  to the banker and the banker gives the player  $\$1$  if a proposition  $A$  is true, and  $0$  otherwise. You do not know if you will be the banker or the player, and you are asked to fix  $p$ . By definition, your fair betting rate  $P(A) = p$  is equated to your personal probability of proposition  $A$ . It is postulated to measure your belief in  $A$ : the more you believe in  $A$ , the more money you will be willing to give to enter the game. Now, the main point is that an opponent can compile a book of bets from your offer that assures a net gain from you (a Dutch book) if and only if  $P$  fails to be a probability function.

To show this, consider two disjoint events  $A$  and  $B$  and the three following bets:

1. Bet 1: the player gains  $\$1$  if  $A$  is true and  $0$  otherwise.
2. Bet 2: the players gain  $\$1$  if  $B$  is true and  $0$  otherwise.
3. Bet 3: the players gain  $\$1$  if  $A \cup B$  is true and  $0$  otherwise.

Let  $P(A)$ ,  $P(B)$  and  $P(A \cup B)$  be the fair prices you are willing to pay for the three tickets. Assume that  $P(A \cup B) < P(A) + P(B)$ . Then, the opponent can raise a Dutch book against you by deciding that you will be the player in the first two bets and the banker in the third bet. You will

---

<sup>1</sup>A Dutch book is a set of odds and bets which guarantees a profit, regardless of the outcome of the gamble.

then have to pay  $P(A) + P(B)$  to participate in the first two bets as a player and you will receive  $P(A \cup B)$  as a banker in the third bet. The balance is thus  $-P(A) - P(B) + P(A \cup B) < 0$ . Now, as shown in Table 1.1, you will not win any additional money, whatever the outcome. For instance, if  $A$  is true and  $B$  is false, you will win \$1 in the first bet but you will lose \$1 in the third bet, and similarly for the two other cases. Hence, you surely incur a net loss. Similarly, if  $P(A \cup B) > P(A) + P(B)$ , you will lose if you are the banker in the first two bets and the player in the third bet. The only way to avoid sure loss is to set the three numbers  $P(A)$ ,  $P(B)$  and  $P(A \cup B)$  such that  $P(A \cup B) = P(A) + P(B)$ .

Table 1.1: Dutch book argument; gains in the three bets.

$A$	$B$	Bet 1	Bet 2	Bet 3
0	0	0	0	0
1	0	1	0	-1
0	1	0	1	-1

### 1.3.3 Operations on probabilities

In the rest of this chapter, we will assume that probabilities represent degrees of belief. However, the operations that will be defined also apply to objective probabilities, sometimes with different justifications. To simplify the exposition, we will also place ourselves in the case where  $\Omega$  is finite, but all the operations defined are easily generalized to the infinite case.

**Bayes' conditioning** Let  $P$  be a probability measure on  $\Omega$  representing your beliefs about some question. Assume that you learn that the truth lies for sure in some subset  $E \subset \Omega$ . How should you update your beliefs to account for this new information? If  $P(E) = 0$ , your previous beliefs are incompatible with the new information (a proposition you believed to be impossible is claimed to be true), and there is no obvious way to reconcile the two pieces of information. If  $P(E) > 0$ , then you can construct a new probability measure  $P^*$  represent your new belief state after receiving the new information. Assume that you impose the following requirements on  $P^*$

1.  $P^*(E) = 1$ .



2. For all  $A, B \subseteq \Omega$ ,  $P^*(B) = 0$  if  $P(B) = 0$ , and

$$\frac{P^*(A)}{P^*(B)} = \frac{P(A)}{P(B)} \quad (1.23)$$

if  $P(B) > 0$ .

The only probability measure verifying these requirements is defined as

$$P^*(A) = P(A|E) = \frac{P(A \cap E)}{P(E)}, \quad (1.24)$$

for all  $A \subseteq \Omega$ . This operation *Bayes' conditioning*. It is the standard mechanism for updating beliefs represented by a probability measure, in light of certain evidence. For any  $A, B \subseteq \Omega$  such that  $P(A) > 0$  and  $P(B) > 0$ , we have

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A), \quad (1.25)$$

which implies that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (1.26)$$

Formula (1.26) is called *Bayes' rule*. It plays a fundamental role in probabilistic reasoning.

If  $P(A \cap B) = P(A)P(B)$ , the events  $A$  and  $B$  are said to be *independent*. It easy to see that, in that case,  $P(A|B) = P(A)$ , and  $P(B|A) = P(B)$ , i.e., knowing that one of the two events is true does not change our assessment of the probability of the other event.

**Jeffrey's conditioning** Jeffrey [48] proposed to extend Bayes' rule as follows. Let  $E_1, \dots, E_n$  be a partition of  $\Omega$ . Assume that you receive some new evidence, and the effect of that evidence does not go beyond changing your degrees of beliefs about  $E_1, \dots, E_n$ . Let  $q_1, \dots, q_n$  be these degrees of belief. We thus want to construct an updated probability measure  $P^*$  such that  $P^*(E_i) = q_i$ ,  $i = 1, \dots, n$ , and

$$P^*(A|E_i) = P(A|E_i), \quad i = 1, \dots, n. \quad (1.27)$$

The unique solution is

$$P^*(A) = \sum_{i=1}^n P(A|E_i)q_i. \quad (1.28)$$

Bayes' rule is recovered as a special case with  $(E_1, E_2) = (E, \bar{E})$ ,  $q_1 = 1$  and  $q_2 = 0$ . This operation is called *Jeffrey's conditioning*. It is a mechanism for updating probabilistic knowledge in light of uncertain evidence.

**Marginalization** As in Section 1.2.1, assume that we have two frames of discernment  $\Omega_X$  and  $\Omega_Y$  for two different questions. Let  $P_{XY}$  be a probability measure on  $\Omega_{XY} = \Omega_X \times \Omega_Y$  with probability mass function  $p_{XY}$ . The marginal probability mass function on  $\Omega_X$  is

$$p_X(x) = P_{XY}(\{x\} \times \Omega_Y) = \sum_{y \in \Omega_Y} p_{X,Y}(x, y). \quad (1.29)$$

This operation is called *marginalization*. It plays the same role as projection (1.6b) in the set-membership setting. If the events  $\{x\} \times \Omega_Y$  and  $\Omega_X \times \{y\}$  are independent for any  $x$  and  $y$ , then

$$p_{X,Y}(x, y) = P_{XY}[(\{x\} \times \Omega_Y) \cap (\Omega_X \times \{y\})] = p_X(x)p_Y(y). \quad (1.30)$$

We then say that the  $X$  and  $Y$  are independent.

### 1.3.4 Justification of probabilities: Cox's theorem

Some scholars have attempted to justify the use of probabilities to represent degrees of belief using an axiomatic approach. In particular, Cox's [14] and Savage's [74] theorems are often invoked by Bayesians to argue that probability theory is the only "reasonable" formalism for reasoning with uncertainty. In this section, we will briefly discuss Cox's theorem (see [102] for a recent discussion of these axioms). Savage's theorem will be discussed in Chapter 7.

Let  $Cr(A|B) \in \mathbb{R}$  be a measure of the "credibility" of proposition  $A$ , given that  $B$  is true, where  $A$  and  $B$  are non-empty subsets of  $\Omega$ . Consider the following axioms:

$A_1$  : The credibility of the complement of  $A$  can be computed from the credibility of  $A$ ,

$$Cr(\bar{A}|B) = S[Cr(A|B)], \quad (1.31)$$

where  $S$  is a twice differentiable function;

$A_2$  : The credibility of  $A \cap A'$  given  $B$  is a function of the credibility of  $A$  given  $B$ , and the credibility of  $A'$ , given  $A \cap B$ ,

$$Cr(A \cap A'|B) = F[Cr(A'|A \cap B), Cr(A|B)], \quad (1.32)$$

where  $F$  is a twice differentiable function with a continuous derivative.

Under these assumptions, Cox showed that  $Cr$  is isomorphic to a probability distribution, in the sense that there exists a one-to-one mapping  $g : \mathbb{R} \rightarrow \mathbb{R}$  such that  $g \circ Cr$  is a probability measure, and

$$g[Cr(A|B)] \cdot g[Cr(B)] = g[Cr(A \cap B)] \quad (1.33)$$

for any  $A$  and non-empty  $B$ , with  $Cr(B) = Cr(B|\Omega)$ .

Significant as it may be, this result can hardly be considered as a final justification of probabilities for representing degrees of belief. Indeed, close inspection of the axioms shows that they can be seriously questioned.

The first assumption is that the credibility of a proposition can be represented by a single number. This condition is not assumed in some alternative theories of uncertainty, such as the theory of belief functions. Axiom  $A_1$  is also quite debatable. If degrees of credibility are identified with degrees of support, the degree of support for some proposition is not a function of the degree of support for its negation (if  $A$  is not supported,  $\bar{A}$  may be supported or not), and  $Cr(A|\Omega)$  will not be determined by  $Cr(\bar{A}|\Omega)$ .

Cox merely justifies axiom  $A_2$  by an example. If  $A$  is the proposition that some athlete can run to some point, given the conditions of the race expressed by  $B$ , and if  $A'$  denotes the proposition that he can come back, then the probability that he can run to the point and come back depends on the probability that he can reach the point, and the probability that he can come back, given that he has already reached the point. Yet, as noted by Shafer [82], even admitting that  $Cr(A \cap A'|B)$  should be a function of  $Cr(A'|A \cap B)$  and  $Cr(A|B)$ , it is not obvious that the same function  $F$  should always be used.

### 1.3.5 Principle of indifference

As we have seen in this chapter, two classical representations of partial knowledge coexist in the literature: sets, and probabilities. Can we translate knowledge from one representation to the other without loss of information? A probability measure can certainly not be transformed into a set without losing information. One could define, for instance, the smaller subset  $A$  of  $\Omega$  such that  $P(A) \geq \alpha$  for some  $\alpha \in (0, 1)$ , but we would merely obtain a summary of the probability measure.

In the other direction, it is often claimed that set-based information can be translated in the language of probabilities using the so-called *Laplace's principle of indifference* (PI). According to this principle, a piece of information of the form  $X \in A$  can be represented by assigning equal probabilities to any possible values of  $X$ . This principle, however, has a major shortcoming:

the probability measure built in that way is very sensitive to the definition of the frame of discernment. For instance, assume that  $A = \{\omega_1, \omega_2\}$ . According to the PI, the probability mass function representing the information  $X \in A$  is  $p(\omega_1) = p(\omega_2) = 1/2$ . However, if possibility  $\omega_2$  is split into two possibilities  $\omega_{21}$  and  $\omega_{22}$ , assigning equal probabilities to the now three possibilities leads to  $p(\omega_1) = p(\omega_2) = p(\omega_3) = 1/3$ . Why should the probability of  $\omega_1$  depend on the degree of detail with which the other possibility  $\omega_2$  is described? A similar problem occurs when  $X$  is a numerical variable and we get information of the form  $X \in [a, b]$ . Following the PI, such information should be represented by a continuous uniform probability measure on  $[a, b]$ . However, if  $f$  is a nonlinear function, the probability distribution of  $f(X)$  will not, in general, be uniform.

Advocates of the PI will object that the frame of discernment and the variables on which the probabilities are assessed are part of our current knowledge about the problem. For instance, one should define the finest frame of discernment compatible with our knowledge. This notion of a “finest frame” can be questioned, however, because a frame can always be refined [77]. Also, there are some cases where the choice of variables describing a problem seems completely arbitrary. The wine/water paradox, attributed to Von Mises, provides a very good illustration of this point (see [61] and [16] for recent reviews and discussions).

**The wine/water paradox** Consider a certain quantity of liquids. All we know is that this liquid is composed entirely of wine and water, and the ratio of wine to water is between  $1/3$  and  $3$ . What is the probability that the ratio of wine to water is less than or equal to  $2$ ?

Let  $X$  denote the ratio of wine to water. All we know is that  $X \in [1/3, 3]$ . According to the PI,  $X \sim \mathcal{U}_{[1/3, 3]}$ . Consequently:

$$P(X \leq 2) = (2 - 1/3)/(3 - 1/3) = 5/8. \quad (1.34)$$

Now, let  $Y = 1/X$  denote the ratio of water to wine. All we know is that  $Y \in [1/3, 3]$ . According to the PI,  $Y \sim \mathcal{U}_{[1/3, 3]}$ . Consequently:

$$P(Y \geq 1/2) = (3 - 1/2)/(3 - 1/3) = 15/16. \quad (1.35)$$

By comparing (1.34) and (1.35), we can see that we have a paradox, as the propositions  $X \leq 2$  and  $Y \geq 1/3$ , being logically equivalent, should receive the same probability.

The reason for this paradox is that, if  $X$  has a uniform distribution on some set  $A$ , and if  $f$  is a non linear mapping,  $f(X)$  does not have, in general,

a uniform distribution on  $f(A)$ . However, if we only know that  $X$  is in  $A$ , we only know that  $f(X)$  is in  $f(A)$ . This argument shows that set-valued information cannot be adequately represented by a probability measure.

## 1.4 Conclusions

The two main formalisms for representing uncertain information are set-based representations and probability theory. We have shown in this chapter that none of these two formalisms seems to be sufficient to represent all kinds of uncertainties. In the next chapter, we will introduce the theory belief functions, which can be seen as generalizing the two classical frameworks outlined above.



Part I

Fundamentals





## Chapter 2

# Representation of Evidence

In this chapter, we define some of the main concepts of Dempster-Shafer theory in the finite case. These notions are sufficient to cope with a large number of applications. The extension to infinite spaces involves some mathematical intricacies and is technically more difficult, except in some simple (and practically important) cases; it is postponed to Chapter 8.

### 2.1 Mass function

#### 2.1.1 Definitions

Let  $\Omega$  be a finite set of possible answers to some question  $Q$ , one and only one of which is true. The true answer will be denoted by  $\omega$ , and an arbitrary element of  $\Omega$  by  $\omega$ . Shafer [77] calls such a space a *frame of discernment*, to emphasize the fact that it is not a set of “states of nature” objectively given, but a subjective construction based on our state of knowledge. For instance, if  $Q$  relates to a person’s state of health,  $\Omega$  might contain only the diseases known at a certain time. This set could be later refined or extended if new knowledge became available. We will come back in Chapter 6 to the important issue of defining and modifying the frame of discernment.

A piece of evidence about  $Q$  will be represented by a *mass function*, defined as a mapping  $m$  from the power set  $2^\Omega$  to the interval  $[0, 1]$ , such that  $m(\emptyset) = 0$  and

$$\sum_{A \subseteq \Omega} m(A) = 1. \quad (2.1)$$

As will be discussed later, each number  $m(A)$  represents the probability that the evidence supports exactly the proposition  $\omega \in A$ , and no more specific

Table 2.1: Four mass functions on  $\Omega = \{a, b, c\}$  in Example 2.1.

$A$	$\emptyset$	$\{a\}$	$\{b\}$	$\{a, b\}$	$\{c\}$	$\{a, c\}$	$\{b, c\}$	$\{a, b, c\}$
$m_1(A)$	0	0.2	0.5	0	0.3	0	0	0
$m_2(A)$	0	0	0	1	0	0	0	0
$m_7(A)$	0	0	0	0	0	0	0	1
$m_3(A)$	0	0.1	0.05	0.2	0.15	0.3	0.1	0.1

proposition. Any subset  $A$  of  $\Omega$  such that  $m(A) > 0$  is called a *focal set* of  $m$ . The union of the focal sets of a mass function is called its *core*.

Before discussing the semantics of a mass function, it is interesting to point out two special cases:

1. If  $m$  has only one focal set, it is said to be *logical*. Logical mass functions are in one-to-one correspondence with subsets of  $\Omega$ : consequently, general mass functions can be viewed as generalized sets. A particular logical mass function plays a special role in the theory; it is the *vacuous* mass function  $m_7$  defined by  $m_7(\Omega) = 1$ ; such a mass function corresponds to a totally uninformative piece of evidence.
2. If all focal sets are singletons (i.e., sets of cardinality one),  $m$  is said to be *Bayesian*. To each Bayesian mass function can be associated a probability distribution  $p : \Omega \rightarrow [0, 1]$  such that  $p(\omega) = m(\{\omega\})$  for all  $\omega \in \Omega$ .

**Example 2.1** Consider the mass functions on  $\Omega = \{a, b, c\}$  shown in Table 2.1. Mass function  $m_1$  is Bayesian,  $m_2$  is logical,  $m_7$  is vacuous, and  $m_3$  has no special form.

A belief function may thus be viewed both as a generalized set and as a non-additive measure. As we will see in Chapters 3 and 6, basic mechanisms for reasoning with belief functions extend both probabilistic operations (such as marginalization and conditioning) and set-theoretic operations (such as projection and intersection).

### 2.1.2 Semantics

The following example will show how the formalism of mass functions can be used to represent a piece of evidence. It will also serve as an illustration of the semantics of mass functions.

**Example 2.2** *A murder has been committed and there are three suspects: Peter, John and Mary. The question  $Q$  of interest is the identity of the murderer and the frame of discernment is  $\Omega = \{\text{Peter, John, Mary}\}$ . The piece of evidence under study is a testimony: a witness saw the murderer. However, this witness is short-sighted and he can only report that he saw a man. Unfortunately, this testimony is also not fully reliable, because we know that the witness is drunk 20 % of the time. How can such a piece of evidence be encoded in the language of mass functions?*

*We can see here that what the testimony tells us about  $Q$  depends on the answer to another question  $Q'$ : Was the witness drunk at the time of the murder? If he was not drunk, we know that the murderer is Peter or John. Otherwise, we know nothing. Since there is 80% chance that the former hypothesis holds, we may assign a 0.8 mass to the set  $\{\text{Peter, John}\}$ , and 0.2 to  $\Omega$ :*

$$m(\{\text{Peter, John}\}) = 0.8, \quad m(\Omega) = 0.2$$

In the above example, we receive a message (a testimony) about  $Q$ , whose meaning depends on the answer to a related question  $Q'$  for which we have a chance model (a probability distribution). We can compare our evidence to a canonical example where we know that the outcomes of a random experiment are  $s_1$  and  $s_2$  with corresponding chances  $p_1 = 0.8$  and  $p_2 = 0.2$ , and the message can only be interpreted with knowledge of the outcome. If the outcome is  $s_1$ , then the meaning is  $\omega \in \{\text{Peter, John}\}$ , otherwise the meaning is  $\omega \in \Omega$ , i.e., the message is totally uninformative.

As remarked by Shafer [79], probability judgements can be made by comparing the available evidence to some canonical example involving a chance setup. In the Bayesian theory (see Section 1.3.2), we compare our evidence to a situation where the truth is governed by chance (e.g., by thinking of the murderer as having been selected at random). In the belief function approach, the canonical example describes a situation where the *meaning of the evidence* is governed by chance.

Two scenarios are specially useful to construct canonical examples for mass functions.

The first scenario involves a machine that has two modes of operation, normal and faulty. We know that in the normal mode it broadcasts true messages, but we are completely unable to predict what it does in the faulty mode. We further assume that the operating mode of the machine is random and there a chance  $p$  that it is in the normal mode. It is then natural to say that a message  $\omega \in A$  produced by the machine has a chance  $p$  of meaning what it says and a chance  $1 - p$  of meaning nothing. This leads to the mass

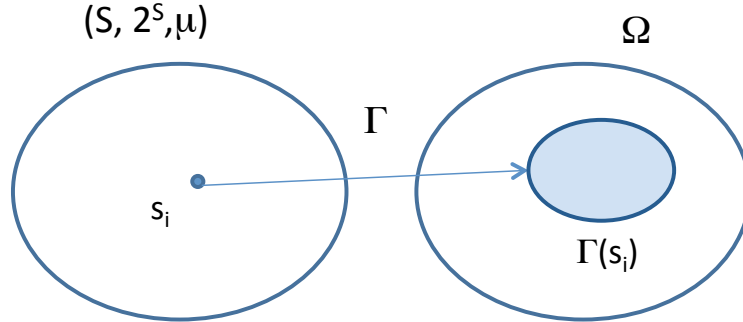


Figure 2.1: Random code setup.

function  $m(A) = p$  and  $m(\Omega) = 1 - p$ . Such a mass function, with two focal sets including  $\Omega$ , is called a simple mass function.

The above story is simple and very useful to model situations in which a partially reliable source of information provides a simple statement of the form  $\omega \in A$  and we can assess the probability of the source to be reliable. How, it is not general enough to cover all kinds of evidence. In [79], Shafer introduced a more sophisticated scenario that is general enough to produce canonical examples for arbitrary mass functions. In this scenario, a source holds some *true information* of the form  $\omega \in A^*$  for some  $A^* \subseteq \Omega$ . It sends us this information as an *encoded message* using a code chosen at random from a set of codes  $S = \{s_1, \dots, s_r\}$ , according to some known probability measure  $\mu$  (Figure 2.1). We know the set of codes as well as the chances of each code to be selected. If we decode the message using code  $s$ , we get a decoded message of the form  $\omega \in \Gamma(s)$  for some subset  $\Gamma(s)$  of  $\Omega$ . Then,

$$m(A) = \mu(\{s \in S \mid \Gamma(s) = A\}) \quad (2.2)$$

is the chance that the original message was “ $\omega \in A$ ”, i.e., the *probability of knowing that  $\omega \in A$* , and nothing more.

In the above framework, the mapping  $\Gamma : S \rightarrow 2^\Omega \setminus \{\emptyset\}$  is called a *multi-valued mapping* and the 4-tuple  $(S, 2^S, \mu, \Gamma)$  is called a *source*. We can observe that a source corresponds formally to a *random set* [67]. However, the term “random set” may be misleading here, because we are not interested in situations where a set is selected at random (such as, e.g., drawing a handful of marbles from a bag). Here, the true answer to the question of interest is a single element of  $\Omega$  and it is not assumed to have been selected at random. Instead, chances are introduced when comparing our evidence

to a situation where the meaning of a message depends on the result of a random experiment.

It is clear that a source  $(S, 2^S, \mu, \Gamma)$  always induces a mass function from (2.2). Conversely, any mass function can be seen as generated by a source. For instance, if  $A_1, \dots, A_n$  are the focal sets of a mass function  $m$ , we may set  $S = \{1, \dots, n\}$  and  $\mu(\{i\}) = m(A_i)$  for  $1 \leq i \leq n$ . However, as we will see in Chapter 8, the concept of a source is more general than that of mass function, because a source can be used in the infinite case to generate a belief function even when a mass function does not exist.

## 2.2 Belief and plausibility functions

### 2.2.1 Definitions

Let us assume the available evidence to be encoded by a mass function  $m$  on  $\Omega$  generated by a source  $(S, 2^S, \mu, \Gamma)$ . For any  $A \subseteq \Omega$ , the uncertainty pertaining to the proposition  $\omega \in A$  can be quantified by two numbers:

1. The probability that the evidence supports (implies)  $A$ , defined by

$$Bel(A) = \mu(\{s \in S | \Gamma(s) \subseteq A\}) \quad (2.3a)$$

$$= \sum_{B \subseteq A} m(B); \quad (2.3b)$$

2. The probability that the evidence does not contradict  $A$ , given by

$$Pl(A) = \mu(\{s \in S | \Gamma(s) \cap A \neq \emptyset\}) \quad (2.4a)$$

$$= \sum_{B \cap A \neq \emptyset} m(B). \quad (2.4b)$$

Clearly,  $Bel(\emptyset) = Pl(\emptyset) = 0$ ,  $Bel(\Omega) = Pl(\Omega) = 1$ ,  $Bel(A) \leq Pl(A)$  and  $Pl(A) = 1 - Bel(\bar{A})$ , where  $\bar{A}$  denotes the complement of  $A$ . The quantity  $Bel(A)$  can be interpreted as a degree of support for proposition  $A$ , or as a degree of belief. The function  $Bel : 2^\Omega \rightarrow [0, 1]$  is called a *belief function*. In contrast,  $Pl(A)$  can be seen as the degree to which one fails to doubt  $A$ ; this number is called the plausibility of  $A$  and the function  $Pl : 2^\Omega \rightarrow [0, 1]$  is called a *plausibility function*.

**Example 2.3** Consider a mass function  $m$  induced by the source shown in Figure 2.2. It has four focal sets  $B_i$ ,  $i = 1, 2, 3, 4$ . The degree of belief

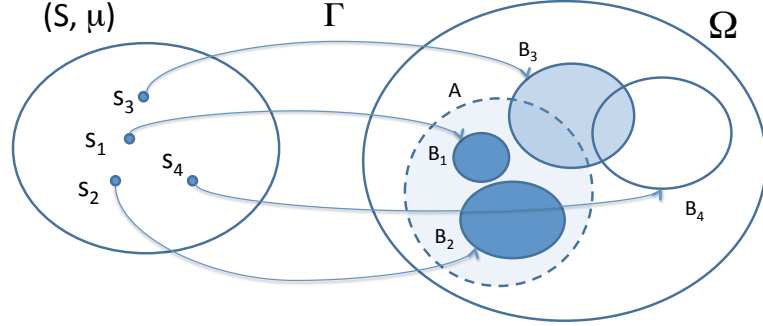


Figure 2.2: Belief and plausibility functions (Example 2.3).

in  $A$  is  $Bel(A) = m(B_1) + m(B_2)$ , while the plausibility of  $A$  is  $Pl(A) = m(B_1) + m(B_2) + m(B_3)$ . The degree of belief in the complement of  $A$  is  $Bel(\bar{A}) = m(B_4)$ , which is clearly equal to  $1 - Pl(A)$ .

## 2.2.2 Properties

**Theorem 2.1** A function  $Bel : 2^\Omega \rightarrow [0, 1]$  is a belief function iff it satisfies the following conditions:

1.  $Bel(\emptyset) = 0$ ;
2.  $Bel(\Omega) = 1$ ;
3. For any  $k \geq 2$  and any collection  $A_1, \dots, A_k$  of subsets of  $\Omega$ ,

$$Bel\left(\bigcup_{i=1}^k A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right). \quad (2.5)$$

Proof: See [77, page 51].

Property (2.5) is a weaker form of the corresponding property (1.20), which holds for probability measure. In general, a function satisfying (2.5) for a given  $k$  is said to be *monotone of order  $k$* . It is clear that monotonicity of order  $k$  implies monotonicity of order  $k'$  for all  $k' < k$ . A function that is monotone for any  $k$  is said to be *monotone of order infinite*, or *completely monotone*. Furthermore, properties 1 and 2 above imply that  $Bel$  is increasing. To see this, let  $A$  and  $B$  be two subsets of  $\Omega$  such that  $A \subseteq B$  and let

$C = B \setminus A$ . We have  $B = A \cup C$  and  $A \cap C = \emptyset$ . From (2.5) with  $k = 2$ , we have

$$\begin{aligned} Bel(B) &= Bel(A \cup C) \geq Bel(A) + Bel(C) - Bel(A \cap C) \\ &= Bel(A) + Bel(C) \geq Bel(A). \end{aligned} \quad (2.6)$$

Theorem 2.1 tells us that a completely monotone set function such that  $Bel(\emptyset) = 0$  and  $Bel(\Omega) = 1$  is induced by some mass function  $m$  using (2.3b). We may wonder whether there exists a unique  $m$  generating a belief function  $Bel$ . Indeed, (2.3b) for  $A \in 2^\Omega \setminus \{\emptyset, \Omega\}$  provides  $2^{|\Omega|} - 2$  equations and there are  $2^{|\Omega|} - 2$  free mass numbers (taking into account constraint (2.1)). Consequently, one must be able to recover  $m$  from  $Bel$  in a unique way. The following theorem states that  $m$  is actually the Möbius inverse of  $Bel$ , a notion from combinatorial theory [77].

**Theorem 2.2** *Let  $Bel : 2^\Omega \rightarrow [0, 1]$  be a belief function induced by a mass function  $m$ . Then*

$$m(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} Bel(B), \quad (2.7)$$

for all  $A \subseteq \Omega$ .

Proof: See [77, page 52].

Using the identity  $Pl(A) = 1 - Bel(\bar{A})$  for any  $A \subseteq \Omega$ , it is easy to obtain the Theorems 2.3 and 2.4, which are counterparts of Theorems 2.1 and 2.2.

**Theorem 2.3** *A function  $Pl : 2^\Omega \rightarrow [0, 1]$  is a plausibility function iff it satisfies the following conditions:*

1.  $Pl(\emptyset) = 0$ ;
2.  $Pl(\Omega) = 1$ ;
3. For any  $k \geq 2$  and any collection  $A_1, \dots, A_k$  of subsets of  $\Omega$ ,

$$Pl\left(\bigcap_{i=1}^k A_i\right) \leq \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} Pl\left(\bigcup_{i \in I} A_i\right). \quad (2.8)$$

A set function verifying (2.8) is said to be *alternating of order infinite*, or *completely alternating*. A plausibility function is thus a completely alternating set function  $Pl$  such that  $Pl(\emptyset) = 0$  and  $Pl(\Omega) = 1$ . Given a plausibility function  $Pl$ , the corresponding mass function  $m$  can be recovered using the following theorem.

**Theorem 2.4** *Let  $Pl : 2^\Omega \rightarrow [0, 1]$  be a plausibility function induced by a mass function  $m$ . Then*

$$m(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|+1} Pl(\bar{B}), \quad (2.9)$$

for all  $A \subseteq \Omega$ .

From the above results, it is clear that, given any of the three functions  $m$ ,  $Bel$  and  $Pl$ , we can recover the other two. Consequently, these three functions can be seen as different facets of the same information. In the sequel, we will sometimes use the term “belief function” to refer to any of these functions, when there will be no risk of confusion.

### 2.2.3 Vector representation

A mass function on a finite frame may be represented as a vector once the subsets of  $\Omega$  have been arranged in some predefined order [91]. One such order with nice properties is the binary order defined as follows. Let  $\Omega = \{\omega_1, \dots, \omega_n\}$  be the frame of discernment. For any  $A \subseteq \Omega$ , let  $u(A)$  be the binary number  $u_n u_{n-1} \dots u_2 u_1$  such that  $u_k$  equals 1 if  $\omega_k \in A$  and 0 otherwise. The rank of  $A$  in the binary ordering is  $r(A) = \sum_{k=1}^n u_k 2^{k-1} + 1$ .

Assuming the focal sets  $A_1, \dots, A_{2^n}$  to be indexed in such a way that  $r(A_i) = i$  for all  $i$ , the vector representation of a mass function  $m$  on  $\Omega$  is the column vector  $\mathbf{m} = (m_1, \dots, m_{2^n})'$ , where  $m_i = m(A_i)$ . This representation is shown in Table 2.2 in the case  $n = 3$ . An interesting property of this ordering is that, whenever  $A_i \subset A_j$ , we always have  $i < j$ .

Table 2.2: Binary ordering in the case  $|\Omega| = 3$ .

$i$	$u(A_i)$	$A_i$	$m_i$
1	000	$\emptyset$	$m(\emptyset)$
2	001	$\{\omega_1\}$	$m(\{\omega_1\})$
3	010	$\{\omega_2\}$	$m(\{\omega_2\})$
4	011	$\{\omega_1, \omega_2\}$	$m(\{\omega_1, \omega_2\})$
5	100	$\{\omega_3\}$	$m(\{\omega_3\})$
6	101	$\{\omega_1, \omega_3\}$	$m(\{\omega_1, \omega_3\})$
7	110	$\{\omega_2, \omega_3\}$	$m(\{\omega_2, \omega_3\})$
8	111	$\{\omega_1, \omega_2, \omega_3\}$	$m(\{\omega_1, \omega_2, \omega_3\})$



Arranging the belief and plausibility numbers in vectors

$$\mathbf{Bel} = (Bel(A_1), \dots, Bel(A_n))'$$

and

$$\mathbf{Pl} = (Pl(A_1), \dots, Pl(A_n))'$$

it is clear that the transformations from any of the three representations  $m$ ,  $Bel$  and  $Pl$  to another are linear. For instance, (2.3b) becomes, in vector notation,

$$\mathbf{Bel} = \mathbf{BfrM} \cdot \mathbf{m}, \quad (2.10)$$

where  $\mathbf{BfrM}$  is a square matrix of size  $2^n$ , whose general term  $BfrM_{ij}$  equals 1 if  $A_i \subseteq A_j$  and 0 otherwise. We can easily check that matrix  $\mathbf{BfrM}$  is lower triangular.

## 2.3 Special cases and related theories

### 2.3.1 Bayesian mass functions

If  $m$  is Bayesian, then

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

for any  $A \subseteq \Omega$ . Furthermore, for any two disjoint subsets  $A$  and  $B$  of  $\Omega$ ,

$$\begin{aligned} Bel(A \cup B) &= \sum_{\omega \in A \cup B} m(\{\omega\}) = \\ &= \sum_{\omega \in A} m(\{\omega\}) + \sum_{\omega \in B} m(\{\omega\}) = Bel(A) + Bel(B). \end{aligned} \quad (2.11)$$

Consequently, belief functions induced by Bayesian mass functions are probability measures and are equal to their dual plausibility functions. Conversely, it is clear that each probability measure  $P$  is a belief function induced by the Bayesian mass function  $m$  such that  $m(\{\omega\}) = P(\{\omega\})$  for all  $\omega \in \Omega$ .

In other terms, the set of probability measures is exactly the set of belief functions induced by Bayesian mass functions. This result shows us that the language of belief functions is more general than that of probability theory. As we will see in Chapter 3, the conditioning operation, which plays a major role in updating beliefs based on new evidence in the Bayesian framework, can also be seen as a special case of a more general operation in the belief function framework.

### 2.3.2 Consonant mass functions

A mass function  $m$  is said to be *consonant* if its focal sets are nested, i.e., if they can be arranged in an increasing sequence  $A_1 \subset \dots \subset A_r$ . In that case, functions  $Bel$  and  $Pl$  satisfy the following properties.

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

$$Bel(A \cap B) = \sum_{i=1}^{\min(i_1, i_2)} m(A_i) \quad (2.12a)$$

$$= \min \left( \sum_{i=1}^{i_1} m(A_i), \sum_{i=1}^{i_2} m(A_i) \right) \quad (2.12b)$$

$$= \min(Bel(A), Bel(B)). \quad (2.12c)$$

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

$$Pl(A \cup B) = 1 - Bel(\overline{A \cup B}) \quad (2.13a)$$

$$= 1 - Bel(\overline{A} \cap \overline{B}) \quad (2.13b)$$

$$= 1 - \min(Bel(\overline{A}), Bel(\overline{B})) \quad (2.13c)$$

$$= \max(1 - Bel(\overline{A}), 1 - Bel(\overline{B})) \quad (2.13d)$$

$$= \max(Pl(A), Pl(B)). \quad (2.13e)$$

Properties (2.12c) and (2.13) characterize, respectively, *possibility* and *necessity* measures, which form the basis of Possibility theory introduced by Zadeh in [117]. In this theory,  $Pl(A)$  is the degree to which proposition  $A$  is possible, and  $Bel(A)$  is the degree to which  $A$  is certain, i.e., the degree to which  $\overline{A}$  is impossible. As possibility measures are special plausibility functions (induced by consonant mass functions), the theory of belief functions can be considered as more expressive than Possibility theory. However, as we shall see in Chapter 3, the two theories depart in the way different pieces of information are combined: in the belief function approach, a mass function resulting from the combination of two consonant mass functions will generally not be consonant.

An important consequence of (2.13) is that function  $Pl$  can be deduced from its restriction to singletons. More precisely, let  $pl : \Omega \rightarrow [0, 1]$  be the *contour function* of  $m$ , defined by  $pl(\omega) = Pl(\{\omega\})$ , for all  $\omega \in \Omega$ . For all  $A \subseteq \Omega$ ,

$$Pl(A) = \max_{\omega \in A} pl(\omega). \quad (2.14)$$

We note that the condition  $Pl(\Omega) = 1$  implies that  $\max_{\omega \in \Omega} pl(\omega) = 1$ . The contour function  $pl$  is then the *possibility distribution* associated to the possibility measure  $Pl$ .

We have seen that the plausibility function induced by a consonant mass function is a possibility measure. Conversely, a possibility measure  $\Pi$  is always a plausibility function for some consonant mass function, which can be recovered from  $\pi$  as explained in the following theorem [27].

**Theorem 2.5** *Let  $\pi$  be a possibility distribution on the frame  $\Omega = \{\omega_1, \dots, \omega_n\}$ , with elements arranged by decreasing order of plausibility, i.e.,*

$$1 = \pi(\omega_1) \geq \pi(\omega_2) \geq \dots \geq \pi(\omega_n),$$

and let  $A_i$  denote the set  $\{\omega_1, \dots, \omega_i\}$ , for  $1 \leq i \leq n$ . Then,  $\pi$  is the contour function for a mass function  $m$  obtained by the following formula:

$$m(A_i) = \pi(\omega_i) - \pi(\omega_{i+1}), \quad 1 \leq i \leq n-1, \quad (2.15a)$$

$$m(\Omega) = \pi(\omega_n). \quad (2.15b)$$

Proof: Let  $A$  be a non empty subset of  $\Omega$ . Let  $\Pi$  be the possibility measure induced by  $\pi$  and let  $Pl$  be the plausibility measure induced by  $m$ , given by (2.15). The possibility of  $A$  is

$$\Pi(A) = \max_{\omega \in A} \pi(\omega) = \pi(\omega_{i_0}) \quad (2.16)$$

for some  $1 \leq i_0 \leq n$ . It is clear that  $A_i$  intersects  $A$  if and only if  $i \geq i_0$ . Consequently,

$$Pl(A) = \sum_{A_i \cap A \neq \emptyset} m(A_i) = \sum_{i=i_0}^n m(A_i) \quad (2.17a)$$

$$= pl(\omega_{i_0}) - pl(\omega_{i_0+1}) + pl(\omega_{i_0+1}) - \dots - pl(\omega_n) \quad (2.17b)$$

$$= pl(\omega_{i_0}) \quad (2.17c)$$

$$= \Pi(A), \quad (2.17d)$$

which completes the proof.  $\square$

**Example 2.4** *Consider, for instance, the following possibility distribution defined on the frame  $\Omega = \{a, b, c, d\}$ :*

$\omega$	$a$	$b$	$c$	$d$
$\pi(\omega)$	0.3	0.5	1	0.7

The corresponding mass function is

$$m(\{c\}) = 1 - 0.7 = 0.3 \quad (2.18a)$$

$$m(\{c, d\}) = 0.7 - 0.5 = 0.2 \quad (2.18b)$$

$$m(\{c, d, b\}) = 0.5 - 0.3 = 0.2 \quad (2.18c)$$

$$m(\{c, d, b, a\}) = 0.3. \quad (2.18d)$$

Possibility theory has a strong connection with the theory of Fuzzy Sets [116]. More precisely, if we receive evidence of the form “ $\omega$  is  $F$ ”, where  $F$  is a fuzzy subset of  $\Omega$  with membership function  $\mu_F$ , then this piece of evidence may be represented by a consonant belief function with contour function  $pl = \mu_F$ .

### 2.3.3 Relation with imprecise probabilities

Let  $\mathcal{P}$  be a non empty set of probability measures on some frame  $\Omega$ . Its *lower and upper envelopes* are set functions defined as follows:

$$P_*(A) = \inf_{P \in \mathcal{P}} P(A), \quad (2.19a)$$

$$P^*(A) = \sup_{P \in \mathcal{P}} P(A). \quad (2.19b)$$

for all subsets  $A$  of  $\Omega$ . Functions  $P_*$  and  $P^*$  are called, respectively, *coherent lower and upper probabilities* [105]. Clearly,

$$P^*(A) = 1 - P_*(\bar{A}) \quad (2.20)$$

for all  $A$ , which is reminiscent of the relation between belief and plausibility functions. What is the relation between these notions?

First of all, we can observe that, to each belief function  $Bel$  we can associate the set of probability measures  $P$  that dominate  $Bel$ , i.e., the set of probability measures such that  $P(A) \geq Bel(A)$  for all subset  $A$  of  $\Omega$ . Because of the relation  $Bel(A) = 1 - Pl(\bar{A})$ , we also have  $P(A) \leq Pl(A)$  for all  $A$ , or

$$Bel(A) \leq P(A) \leq Pl(A), \quad \forall A \subseteq \Omega. \quad (2.21)$$

Any probability measure  $P$  verifying (2.21) is said to be *compatible* with  $Bel$ , and the set  $\mathcal{P}(Bel)$  of all probability measures compatible with  $Bel$  is called the *credal set* of  $Bel$ .

An arbitrary element of  $\mathcal{P}(Bel)$  can be obtained by distributing each mass  $m(A)$  among the elements of  $A$ . More precisely, let us call an *allocation* of  $m$  any function

$$\alpha : \Omega \times 2^\Omega \setminus \{\emptyset\} \rightarrow [0, 1] \quad (2.22)$$

such that, for all  $A \subseteq \Omega$ ,

$$\sum_{\omega \in A} \alpha(\omega, A) = m(A). \quad (2.23)$$

Each quantity  $\alpha(\omega, A)$  can be viewed as a part of  $m(A)$  allocated to the element  $\omega$  of  $A$ . By summing up the numbers  $\alpha(\omega, A)$  for each  $\omega$ , we get a probability mass function on  $\Omega$ ,

$$p_\alpha(\omega) = \sum_{A \ni \omega} \alpha(\omega, A). \quad (2.24)$$

It can be shown [18] that the set of probability measures constructed in that way is exactly equal to the credal set  $\mathcal{P}(Bel)$ . Furthermore, the bounds in (2.21) are attained. A belief function is thus a coherent lower probability. However, a coherent lower probability is not always a belief function. To see this, consider, for instance, the following counterexample taken from [105, page 274]. Suppose a fair coin is tossed twice, in such a way that the outcome of the second toss may depend on the outcome of the first toss. The outcome of the experiment can be denoted by  $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$ . Let  $H_1 = \{(H, H), (H, T)\}$  be the event that we get a head in the first toss,  $H_2 = \{(H, H), (T, H)\}$  the event that we get a head in the second toss, and let  $\mathcal{P}$  be the set of probability measures on  $\Omega$  which assign  $P(H_1) = P(H_2) = 1/2$  and have an arbitrary degree of dependence between tosses. Let  $P_*$  be the lower envelope of  $\mathcal{P}$ . It is clear that  $P_*(H_1) = 1/2$ ,  $P_*(H_2) = 1/2$  and  $P_*(H_1 \cap H_2) = 0$  (as the occurrence of a head in the first toss may never lead to a head in the second toss). Now, in the case of complete positive dependence,  $P(H_1 \cup H_2) = P(H_1) = 1/2$ , hence  $P_*(H_1 \cup H_2) \leq 1/2$ . We thus have

$$P_*(H_1 \cup H_2) < P_*(H_1) + P_*(H_2) - P_*(H_1 \cap H_2), \quad (2.25)$$

which violates the complete monotonicity condition (2.5) for  $k = 2$ .

Mathematically, the notion of coherent lower probability is thus more general than that of belief function. However, the definition of the credal set associated with a belief function is purely formal, as these probabilities have no particular interpretation in our framework. The theory of belief functions is not a theory of imprecise probabilities.

## Exercises

1. Let  $Bel$  be a belief function on  $\Omega$  and let  $Pl$  be the corresponding plausibility function. Show directly (without using Theorems 2.1 or

2.3) that

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

and

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

for all  $A, B \subseteq \Omega$ .

2. Let  $m$  be the mass function on  $\Omega = \{a, b, c\}$  defined by:

$$m(\{a\}) = 0.2 \quad m(\{a, b\}) = 0.5 \quad m(\Omega) = 0.3.$$

Compute  $Bel(A)$  and  $Pl(A)$  for all  $A \subseteq \Omega$ . Which special properties do these functions verify?

3. Represent the uncertainty about the outcome of the Ellsberg's experiment described in Section 1.3.5, using a mass function on a suitable frame. Compute the corresponding belief and plausibility functions.
4. Let us consider the following plausibility function on  $\Omega = \{a, b, c\}$ :

$A$	$\emptyset$	$\{a\}$	$\{b\}$	$\{a, b\}$	$\{c\}$	$\{a, c\}$	$\{b, c\}$	$\{a, b, c\}$
$Pl(A)$	0	0.5	0.55	1	0.5	0.8	0.7	1

Compute the corresponding mass function.

5. Let  $\pi$  be the following possibility distribution on  $\Omega = \{a, b, c, d, e, f\}$ :

$\omega$	$a$	$b$	$c$	$d$	$e$	$f$
$\pi(\omega)$	0.1	0.3	0.5	1	0.7	0.3

Compute the corresponding mass function.

6. Let  $m$  be a consonant mass function on a frame  $\Omega$  and let  $Bel$  and  $Pl$  be the corresponding belief and plausibility functions. Show that, for any subset  $A$  of  $\Omega$ ,  $Bel(A) > 0 \Rightarrow Pl(A) = 1$ .

## Chapter 3

# Combination of independent items of evidence

As discussed in Chapter 2, the theory of belief functions essentially models the process whereby degrees of belief are constructed from pieces of evidence. As several pieces of evidence are typically available, we need a mechanism for combining them. This issue will be addressed in this chapter.

### 3.1 Introductory example

Let us come back to the murder story of Example 2.2. Remember that the first item of evidence gave us the following mass function

$$m_1(\{\text{Peter, John}\}) = 0.8, \quad m_1(\Omega) = 0.2$$

over the frame  $\Omega = \{\text{Peter, John, Mary}\}$ . Let us now assume that we have a new piece of evidence: a blond hair has been found. This new evidence supports the hypothesis that the murderer is either John or Mary, as they are blond while Peter is not. However, this piece of evidence is reliable only if the room has been cleaned before the crime. If we judge that there is 60% chance that it is the case, then our second piece of evidence can be modeled by the following mass function :  $m_2(\{\text{John, Mary}\}) = 0.6$ ,  $m_2(\Omega) = 0.4$ .

The process for combining these two pieces of evidence is illustrated by Figure 3.1. The meaning of each piece of evidence depends on the answer to some related question, which can be seen as being generated by a random process with known chances. For instance, if the witness was not drunk, we know that the murderer is either Peter or John. If the room had been cleaned before the crime, we know that the murderer was either John or Mary. If

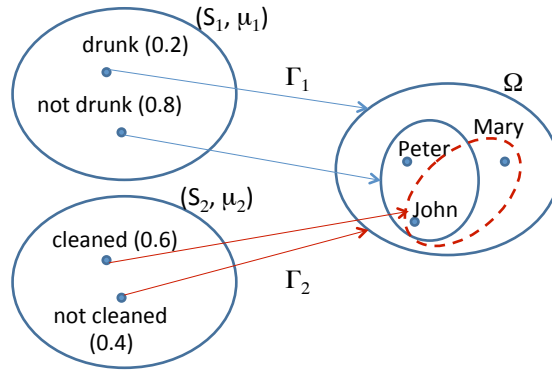


Figure 3.1: Combination of evidence in the murder example.

both assumptions hold, then we know that the murderer is John. What is the probability that this conclusion can be derived from the available evidence? To answer this question, we need to describe the dependence between the two pieces of evidence by specifying a joint probability measure  $\mu_{12}$  on the product space  $S_1 \times S_2$ . Independence between the two pieces of evidence corresponds to the case where  $\mu_{12}$  is the product measure  $\mu_1 \otimes \mu_2$ . Under this independence assumption, the probability of knowing that the murder is John is equal to  $0.6 \times 0.8 = 0.48$ . As the product space  $S_1 \times S_2$  has four elements, there are four cases to consider, which can be summarized in the following table, where, for each case  $(s_1, s_2)$ , we give the set  $\Gamma_1(s_1) \cap \Gamma_2(s_2)$  and the corresponding probability  $\mu_1(\{s_1\})\mu_2(\{s_2\})$ :

	cleaned	not cleaned
drunk	{John, Mary}, 0.12	$\Omega$ , 0.08
not drunk	{John}, 0.48	{Peter, John}, 0.32

We then get the following combined mass function,

$$\begin{aligned}
 m(\{\text{John, Mary}\}) &= 0.12, & m(\Omega) &= 0.08 \\
 m(\{\text{John}\}) &= 0.48, & m(\{\text{Peter, John}\}) &= 0.32.
 \end{aligned}
 \tag{3.1}$$

In some cases, there may be some conflict between two pieces of evidence being combined. For instance, suppose now that only Mary is blond. If we assume that the witness was not drunk and the room had been cleaned before the crime, we get a logical contradiction. Consequently, these two interpretations cannot hold jointly and the joint probability measure on  $S_1 \times$



$S_2$  must be conditioned to eliminate this as well as other conflicting pairs of interpretations. In our second example, we start from the following table:

	cleaned	not cleaned
drunk	{Mary}, 0.12	$\Omega$ , 0.08
not drunk	$\emptyset$ , 0.48	{Peter, John}, 0.32

After conditioning to eliminate the pair (not drunk, cleaned), we get

	cleaned	not cleaned
drunk	{Mary}, 0.12/0.52	$\Omega$ , 0.08/0.52
not drunk	$\emptyset$ , 0	{Peter, John}, 0.32/0.52

which yields the following combined mass function,

$$\begin{aligned} m(\{\text{Mary}\}) &= 0.12/0.52, & m(\Omega) &= 0.08/0.52 \\ m(\emptyset) &= 0, & m(\{\text{Peter, John}\}) &= 0.32. \end{aligned} \quad (3.2)$$

It is clear that such conditioning induces some dependence between the two pieces of evidence. For instance, in the second version of the story, if we learn that the room had been cleaned, then we can deduce that the witness was drunk at the time of the crime. This fact seems to be contradictory with our initial claim that the two pieces of evidence are independent. However, this apparent contradiction is resolved if we consider the meanings of the two pieces of evidence to be governed by a physical chance process, as in the random code metaphor. If  $S_1$  and  $S_2$  are seen as sets of codes selected at random, then independence of the two pieces of evidence corresponds to the assumption of stochastic independence of the two random experiments. After these experiments have taken place, we know that pairs of codes  $(s_1, s_2)$  in  $S_1 \times S_2$  such that  $\Gamma_1(s_1) \cap \Gamma_2(s_2) = \emptyset$  could not have been selected and we must condition  $\mu_1 \otimes \mu_2$  on the event  $\{(s_1, s_2) \in S_1 \times S_2 \mid \Gamma_1(s_1) \cap \Gamma_2(s_2) \neq \emptyset\}$ . This line of reasoning leads to Dempster's rule for combining mass functions, which will be formally defined in the next section.

## 3.2 Dempster's rule

### 3.2.1 Definition and elementary properties

Let  $\mathcal{M}$  be the set of mass functions on  $\Omega$ . Dempster's rule is the partial binary operation  $\oplus$  on  $\mathcal{M}$  defined by

$$(m_1 \oplus m_2)(A) = K \sum_{B \cap C = A} m_1(B)m_2(C) \quad (3.3a)$$

for all  $A \subseteq \Omega$ ,  $A \neq \emptyset$  and

$$(m_1 \oplus m_2)(\emptyset) = 0. \tag{3.3b}$$

This operation is also called the *orthogonal sum*. The normalizing constant  $K$  in (3.3a) is equal to  $(1 - \kappa)^{-1}$ , where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \tag{3.4}$$

is called the *degree of conflict* between  $m_1$  and  $m_2$ . The two mass functions can be combined only if  $\kappa < 1$ , which is the reason why  $\oplus$  is a *partial* binary operation.

**Example 3.1** Consider, for example, the frame  $\Omega = \{a, b, c\}$  and the following mass functions,

$A$	$\emptyset$	$\{a\}$	$\{b\}$	$\{a, b\}$	$\{c\}$	$\{a, c\}$	$\{b, c\}$	$\{a, b, c\}$
$m_1(A)$	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

To combine  $m_1$  and  $m_2$ , it is convenient to present the calculations in a table in which each rows corresponds to a focal set  $B$  of  $m_1$  and each column corresponds to a focal set  $C$  of  $m_2$ . The corresponding cell contains  $B \cap C$  with the mass  $m_1(B)m_2(C)$ . Here, we have

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

The degree of conflict is

$$\kappa = 0.05 + 0.25 + 0.1 = 0.4 \tag{3.5}$$

and the combined mass function is

$$(m_1 \oplus m_2)(\{a\}) = (0.02 + 0.03 + 0.12)/0.6 = 0.17/0.6 \tag{3.6a}$$

$$(m_1 \oplus m_2)(\{b\}) = 0.2/0.6 \tag{3.6b}$$

$$(m_1 \oplus m_2)(\{a, b\}) = 0.08/0.6 \tag{3.6c}$$

$$(m_1 \oplus m_2)(\{c\}) = 0.15/0.6. \tag{3.6d}$$

We may observe that each focal set of  $m_1 \oplus m_2$  is obtained by intersecting one focal set of  $m_1$  and one focal set of  $m_2$ . Consequently,  $m_1 \oplus m_2$  is more focussed (precise) than both  $m_1$  and  $m_2$ : we say that  $\oplus$  is a *conjunctive* operation. Two special cases are of particular interest:

1. If  $m_A$  and  $m_B$  are logical mass functions focussed, respectively, on  $A$  and  $B$  and if  $A \cap B \neq \emptyset$ , then they are combinable and  $m_A \oplus m_B = m_{A \cap B}$ : Dempster's rule thus extends set intersection.
2. 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).

It is clear from (3.3) that  $\oplus$  is commutative ( $m_1 \oplus m_2 = m_2 \oplus m_1$  for any  $m_1$  and  $m_2$ ) and that it admits the vacuous mass function  $m_?$  as neutral element ( $m \oplus m_? = m_? \oplus m = m$  for any  $m$ ). We may wonder whether  $\oplus$  is associative, i.e., for any three mass functions  $m_1$ ,  $m_2$  and  $m_3$ , do we have  $(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$ ? In other words, does the order in which the mass functions are combined matter? Actually, it does. This property will become obvious once Dempster's rule is expressed in terms of another representation of a mass functions: the commonality function introduced in the next section.

### 3.2.2 Commonality function

We have already encountered in Sections 2.1 and 2.2 three equivalent representations of a piece of evidence: the mass function  $m$ , the belief function  $Bel$  and the plausibility function  $Pl$ . There actually exists a fourth representation: the commonality function defined by

$$Q(A) = \sum_{B \supseteq A} m(B), \quad (3.7)$$

for all  $A \subseteq \Omega$ . It can be shown (see [77, Chapter 2]) that  $m$ ,  $Bel$  and  $Pl$  can be uniquely recovered from  $Q$  using the following equations:

$$m(A) = \sum_{B \supseteq A} (-1)^{|B|-|A|} Q(B) \quad (3.8a)$$

$$Bel(A) = \sum_{B \subseteq \bar{A}} (-1)^{|B|} Q(B), \quad (3.8b)$$

$$Pl(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|B|+1} Q(B), \quad (3.8c)$$

for all  $A \subseteq \Omega$ . Conversely,  $Q$  can be directly computed from  $Bel$  or  $Pl$  as follows:

$$Q(A) = \sum_{B \subseteq A} (-1)^{|B|} Bel(\overline{B}), \quad (3.9a)$$

$$Q(A) = \sum_{B \subseteq A} (-1)^{|B|+1} Pl(B), \quad (3.9b)$$

for all  $A \subseteq \Omega$ .

It is obvious that  $Q(\emptyset) = 1$ . Furthermore, using (3.8a) or (3.8b) with  $A = \emptyset$ , we get

$$\sum_{B \subseteq \Omega} (-1)^{|B|} Q(B) = 0 \quad (3.10)$$

or, equivalently,

$$\sum_{\emptyset \neq B \subseteq \Omega} (-1)^{|B|+1} Q(B) = 1. \quad (3.11)$$

Equation (3.11) makes it possible to compute the commonality function once commonality numbers are determined up to some multiplicative constant.

The interpretation of the commonality function is not as obvious as that of the belief and plausibility functions. However, it has a remarkable property in relation with Dempster's rule, as described by the following theorem.

**Theorem 3.1** *Let  $Q_1$ ,  $Q_2$  and  $Q_1 \oplus Q_2$  be the commonality functions induced by mass functions  $m_1$ ,  $m_2$  and  $m_1 \oplus m_2$ . Then*

$$(Q_1 \oplus Q_2)(A) = K Q_1(A) \cdot Q_2(A), \quad (3.12)$$

for all  $A \subseteq \Omega$ ,  $A \neq \emptyset$ , where  $K$  is the same constant as in (3.3a).

*Proof.* For any subset  $A$  of  $\Omega$ , we have

$$\begin{aligned}
(Q_1 \oplus Q_2)(A) &= \sum_{B \supseteq A} (m_1 \oplus m_2)(B) \\
&= K \sum_{B \supseteq A} \sum_{C \cap D = B} m_1(C) m_2(D) \\
&= K \sum_{C \cap D \supseteq A} m_1(C) m_2(D) \\
&= K \sum_{C \supseteq A, D \supseteq A} m_1(C) m_2(D) \\
&= K \left( \sum_{C \supseteq A} m_1(C) \right) \left( \sum_{D \supseteq A} m_2(D) \right) \\
&= K Q_1(A) \cdot Q_2(A).
\end{aligned}$$

Given two mass functions  $m_1$  and  $m_2$ , we can thus combine them either using (3.3), or by converting them to commonality functions, multiplying them pointwise, and computing the corresponding mass function using (3.8a).

Let us now assume that we wish to combine  $n$  mass functions  $m_1, \dots, m_n$ . It can be done by combining  $m_1$  with  $m_2$ , then combining the result  $m_1 \oplus m_2$  with  $m_3$ , etc. The resulting commonality function after combining the  $n$  mass functions is

$$Q(A) = K Q_1(A) \dots Q_n(A) \quad (3.13)$$

for all non-empty  $A \subseteq \Omega$ , where  $K$  is the product of normalizing constants obtained at each stage. Using (3.11), we get the expression of  $K$  as:

$$K = \left( \sum_{\emptyset \neq B \subseteq \Omega} (-1)^{|B|+1} Q_1(B) \dots Q_n(B) \right)^{-1}. \quad (3.14)$$

As both (3.13) and (3.14) are unaffected by permutation of indices, we can conclude that  $\oplus$  is associative and the result of the combination does not depend on the order in which the combination is performed. We can remark that  $m$  can also be computed directly by combining the  $n$  mass functions  $m_1, \dots, m_n$  at once using the following formula, which extends (3.3):

$$(m_1 \oplus \dots \oplus m_n)(A) = K \sum_{B_1 \cap \dots \cap B_n = A} m_1(B_1) \dots m_n(B_n) \quad (3.15a)$$

for all  $A \subseteq \Omega$ ,  $A \neq \emptyset$  and

$$(m_1 \oplus \dots \oplus m_n)(\emptyset) = 0, \quad (3.15b)$$

with  $K = (1 - \kappa)^{-1}$  and

$$\kappa = \sum_{B_1 \cap \dots \cap B_n = \emptyset} m_1(B_1) \dots m_n(B_n). \quad (3.16)$$

As mentioned above,  $\kappa$  is called the degree of conflict between the  $n$  mass function. It ranges between 0 (no conflict) to 1 (total conflict). A related, and perhaps more useful notion is that of *weight of conflict* [77], defined as

$$\text{Con}(m_1, \dots, m_n) = \log K = -\log(1 - \kappa). \quad (3.17)$$

As the normalizing constant  $K$  obtained when combining  $n$  mass functions is equal to the product of the normalizing constants at each stage, it follows that the weights of conflict combine additively, i.e.,

$$\text{Con}(m_1, \dots, m_{n+1}) = \text{Con}(m_1, \dots, m_n) + \text{Con}(m_1 \oplus \dots \oplus m_n, m_{n+1}). \quad (3.18)$$

Theorem 3.1 also has an interesting implication in term of decision-making (see Chapter 7). Assume that our goal when combining  $n$  mass functions is only to compute the plausibility of each element of  $\Omega$ , in view, e.g., of selecting the most plausible element (see Chapter 7). Then, by noticing that  $pl(\omega) = Q(\{\omega\})$  for any  $\omega \in \Omega$ , we can obtain the contour function  $pl$  of  $m_1 \oplus \dots \oplus m_n$ , up to a multiplicative constant, as the product of the contour functions of the  $m_i$ 's:

$$pl = Kpl_1 \dots pl_n, \quad (3.19)$$

which does not require to compute the whole combined mass function.

### 3.2.3 Conditioning

In Bayesian probability theory, conditioning is the fundamental mechanism for updating a probability measure  $P$  with new evidence of the form  $\omega \in B$  for some  $B \subseteq \Omega$  such that  $P(B) \neq 0$ . The conditional probability measure is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (3.20)$$

for all  $A \subseteq \Omega$ . In a similar way, a conditioning rule for mass functions can be defined as a special case of Dempster's rule, in which an arbitrary mass function  $m$  is combined with a logical mass function  $m_B$  focused on  $B$ :

$$m(\cdot|B) = m \oplus m_B. \quad (3.21)$$

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

$$m(A|B) = K \sum_{C \cap B = A} m(C), \quad (3.22)$$

where the normalizing constant  $K$  is

$$K = \left( \sum_{C \cap B \neq \emptyset} m(C) \right)^{-1} = Pl(B)^{-1}.$$

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

$$Pl(A|B) = \sum_{C: C \cap A \neq \emptyset} m(C|B) \quad (3.23a)$$

$$= Pl(B)^{-1} \sum_{C: C \cap A \neq \emptyset} \sum_{D: D \cap B = C} m(D) \quad (3.23b)$$

$$= Pl(B)^{-1} \sum_{D: D \cap B \cap A \neq \emptyset} m(D) \quad (3.23c)$$

$$= \frac{Pl(A \cap B)}{Pl(B)}. \quad (3.23d)$$

We note the similarity between (3.20) and (3.23d). In particular, if  $m$  is Bayesian,  $Pl$  is a probability measure, and  $Pl(\cdot|B)$  is the conditional probability measure obtained by the Bayesian conditioning of  $Pl$  by  $B$ . This fact implies that Dempster's rule can be seen as a proper extension of Bayesian conditioning, which is nothing but Dempster's combination of a probability measure with a logical mass function.

The expression of the conditional belief function  $Bel(\cdot|B)$  can easily ob-

tained from  $Pl(\cdot|B)$ . We have

$$Bel(A|B) = 1 - Pl(\bar{A}|B) \quad (3.24a)$$

$$= 1 - \frac{Pl(\bar{A} \cap B)}{Pl(B)} \quad (3.24b)$$

$$= 1 - \frac{1 - Bel(A \cup \bar{B})}{1 - Bel(\bar{B})} \quad (3.24c)$$

$$= \frac{Bel(A \cup \bar{B}) - Bel(\bar{B})}{1 - Bel(\bar{B})}. \quad (3.24d)$$

### 3.2.4 Computational complexity

The orthogonal sum of two mass functions  $m_1$  and  $m_2$  can be performed in two ways: either directly using (3.3), or by computing the product of commonalities. Using the former, mass-based approach, the time needed to compute the combination is proportional to  $|\mathcal{F}(m_1)||\mathcal{F}(m_2)||\Omega|$ , where  $\mathcal{F}(m_i) \subseteq 2^\Omega$  is the collection of focal sets of  $m_i$  [107]. In the worst case where both mass functions have  $2^{|\Omega|} - 1$  focal sets, the computing time thus becomes proportional to  $2^{2|\Omega|}|\Omega|$ . The other approach implies converting the mass functions into commonalities using (3.7), multiplying the commonalities pointwise, normalizing, and computing the combined mass function using (3.8a). The conversion from one of the equivalent functions  $m$ ,  $Bel$ ,  $Pl$  and  $Q$  to another can be performed using the Fast Möbius transform [52], which takes time proportional to  $|\Omega|2^{2|\Omega|}$ . The mass-based approach is thus more efficient when the number of focal sets is much smaller than the cardinality of  $2^\Omega$ .

Although the combination of mass functions has, in the worst case, exponential complexity, this is rarely an obstacle for practical applications of Dempster-Shafer theory, for several reasons. First, elementary mass functions to be combined often have a simple form, which can considerably simplify the calculations. For instance, the combination of simple mass functions (see Section 3.4) with focal sets of the form  $\{\omega\}$  or  $\{\bar{\omega}\}$  can be performed in time proportional to the size of the frame [8].

A second reason why complexity is usually not prohibitive is that the ultimate goal of uncertain reasoning is often to make decisions. As mentioned in Section 3.2.2, if one seeks the element of  $\Omega$  with the largest plausibility, then we do not need to compute the whole combined mass function. Again, computing the combined contour function can be done in time proportional to the size of the frame. This important property makes it possible to apply



Dempster-Shafer reasoning in huge frame of discernment (such as, e.g., the set of all partitions of of dataset).

Finally, if computing time is limited, we may resort to approximations. As the computational complexity of the mass-based algorithm depends heavily on the number of focal sets, a useful strategy may be to approximate each mass function by a simpler mass function with fewer focal sets. Several methods with different degrees of complexity have been proposed for this purpose [58, 100, 9, 40, 23]. The simplest, yet quite effective approach, is the Summarization algorithm [58], which works as follows. Let  $F_1, \dots, F_n$  be the focal sets of  $m$  ranked by decreasing mass, i.e.,  $m(F_1) \geq m(F_2) \geq \dots \geq m(F_n)$ . If  $n$  exceeds some the maximum allowed number  $k$  of focal sets, then the  $n - k$  focal sets  $F_i$ ,  $i = k + 1, \dots, n$  with the smallest masses are replaced by their union, and  $m$  is approximated by the mass function  $m'$  defined as

$$m'(F_i) = m(F_i), \quad i = 1, \dots, k, \quad (3.25a)$$

$$m' \left( \bigcup_{i=k+1}^n F_i \right) = \sum_{i=k+1}^n m(F_i). \quad (3.25b)$$

As we will see in Chapter 4, mass function  $m'$  can be considered to be less precise, or less “committed” than  $m$ . Using a completely different approach, the combination of several belief functions can also be performed by Monte-Carlo simulation (see, e.g., [65]).

### 3.3 Related combination rules

Let  $(S_1, 2^{S_1}, \mu_1, \Gamma_1)$  and  $(S_2, 2^{S_2}, \mu_2, \Gamma_2)$  be two sources generating mass functions  $m_1$  and  $m_2$ . The combined mass function  $m_1 \oplus m_2$  is induced by the source  $(S_1 \times S_2, 2^{S_1 \times S_2}, \mu, \Gamma_\cap)$ , where  $\mu$  is obtained by conditioning  $\mu_1 \otimes \mu_2$  with the event  $\{(s_1, s_2) \in S_1 \times S_2 | \Gamma_\cap(s_1, s_2) \neq \emptyset\}$ .

When deriving Dempster’s rule, we have made two important assumptions. First, we have assumed both sources to be reliable. In the random code metaphor, this corresponds to the hypothesis that each source encodes a message containing some true information about  $\omega$ . This assumption is at the origin of selecting  $\Gamma_\cap$  as the multi-valued mapping for the combined mass function. We could, however, make different assumptions about the reliability of the two sources. For instance, we could assume that *at least one of them is reliable* [94]. In that case, assuming the codes  $s_1$  and  $s_2$  to be used, we can deduce that  $\omega \in \Gamma_\cup(s_1, s_2) = \Gamma_1(s_1) \cup \Gamma_2(s_2)$ . This assumption results in the following binary operation, called the *disjunctive rule of*

combination [29, 89]:

$$(m_1 \cup m_2)(A) = \sum_{B \cup C = A} m_1(B)m_2(C), \quad (3.26)$$

for all  $A \subseteq \Omega$ .

**Example 3.2** *Let us consider again the two mass functions of Example 3.1. To combine them using the disjunctive rule, we may present the calculations as in the following table,*

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

The resulting mass function is

$$(m_1 \cup m_2)(\{a, b\}) = 0.05 + 0.2 + 0.02 + 0.08 = 0.35 \quad (3.27a)$$

$$(m_1 \cup m_2)(\{b, c\}) = 0.25 \quad (3.27b)$$

$$(m_1 \cup m_2)(\{a, c\}) = 0.03 + 0.15 = 0.18 \quad (3.27c)$$

$$(m_1 \cup m_2)(\Omega) = 0.1 + 0.12 = 0.22. \quad (3.27d)$$

This operation is clearly commutative and associative, and it does not have a neutral element. We can observe that it never generates conflict, so that no normalization has to be performed. The disjunctive rule can be expressed in a simple way using belief functions: if  $Bel_1 \cup Bel_2$  denotes the belief function corresponding to  $m_1 \cup m_2$ , we have

$$(Bel_1 \cup Bel_2)(A) = Bel_1(A)Bel_2(A), \quad (3.28)$$

for all  $A \subseteq \Omega$ , which is the counterpart of (3.12). Combining mass functions disjunctively can be seen as a conservative strategy, as the disjunctive rule relies on a weaker assumption about the reliability of the sources, as compared to Dempster's rule. However, mass functions become less and less focussed as more pieces of information are combined using the disjunctive rule. In particular, the vacuous mass function  $m_?$  is an absorbing element, i.e.,  $m \cup m_? = m_? \cup m = m_?$  for all  $m$ .

In general, the disjunctive rule may be preferred in case of heavy conflict between the different pieces of evidence. An alternative rule, which is somehow intermediate between the disjunctive and conjunctive rules, has been

proposed by Dubois and Prade [29]. It is defined as follows:

$$(m_1 \uplus m_2)(A) = \sum_{B \cap C = A} m_1(B)m_2(C) + \sum_{\{B \cap C = \emptyset, B \cup C = A\}} m_1(B)m_2(C), \quad (3.29)$$

for all  $A \subseteq \Omega$ ,  $A \neq \emptyset$ , and  $(m_1 \uplus m_2)(\emptyset) = 0$ . This rule boils down to the conjunctive and disjunctive rules when, respectively, the degree of conflict is equal to zero and one. In other cases, it has some intermediate behavior. We note that this rule is not associative. If several pieces of evidence are available, they should be combined at once using an obvious  $n$ -ary extension of (3.29).

**Example 3.3** *Let us consider once more the two mass functions of Examples 3.1 and 3.2. To intermediate calculations to combine them using the Dubois-Prade rule are given in the following table,*

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

The resulting mass function is

$$(m_1 \uplus m_2)(\{a, b\}) = 0.05 + 0.08 = 0.13 \quad (3.30a)$$

$$(m_1 \uplus m_2)(\{b\}) = 0.2 \quad (3.30b)$$

$$(m_1 \uplus m_2)(\{b, c\}) = 0.25 \quad (3.30c)$$

$$(m_1 \uplus m_2)(\{a\}) = 0.02 + 0.03 + 0.12 = 0.17 \quad (3.30d)$$

$$(m_1 \uplus m_2)(\{c\}) = 0.15 \quad (3.30e)$$

$$(m_1 \uplus m_2)(\Omega) = 0.1. \quad (3.30f)$$

The other fundamental assumption underlying Dempster's rule is independence of the sources of evidence, which is at the origin of the selection of  $\mu_1 \otimes \mu_2$  as a joint probability measure on  $S_1 \times S_2$ . In principle, any form of dependence between the two sources can be described by defining a joint probability measure  $\mu_{12}$  on  $S_1 \times S_2$ , with marginals  $\mu_1$  and  $\mu_2$ . To each joint measure  $\mu_{12}$  corresponds a distinct combination rule. In practice, however, the dependence between two sources can rarely be specified in that way. Another situation is that where the dependence between sources is unknown. In that case, we could try to find a minimally informative joint probability measure  $\mu_{12}^*$ , among all joints measures with marginals  $\mu_1$  and  $\mu_2$ . This is still a research problem. We will get back to it in Section 4.3.4.

### 3.4 Separable belief functions

Dempster's rule provides the fundamental mechanism for combining elementary items of evidence. The simplest form of such evidence corresponds to the situation where we get a message from a source of the form  $\omega \in A$  for some non-empty  $A \subset \Omega$ , and we assess the chance for the source to be reliable is  $p$ . Such evidence can be represented by a *simple mass function* of the form

$$m(A) = p, \quad m(\Omega) = 1 - p.$$

Shafer [77] defined the weight of evidence associated to  $m$  as  $w = -\log(1-p)$ . The weight of evidence thus equals 0 if  $m$  is vacuous, and  $\infty$  if  $m$  is logical. The interest of the notion of weight of evidence arises from the following observation.

Let  $m_1$  and  $m_2$  be two simple mass functions with the same focal set  $A \subset \Omega$  and masses  $p_1$  and  $p_2$ . Then  $m_1 \oplus m_2$  is still a simple mass function and it is given by

$$(m_1 \oplus m_2)(A) = 1 - (1 - p_1)(1 - p_2) \quad (3.31)$$

$$(m_1 \oplus m_2)(\Omega) = (1 - p_1)(1 - p_2). \quad (3.32)$$

The weight of evidence associated to  $m_1 \oplus m_2$  is thus

$$w_{12} = -\log((1 - p_1)(1 - p_2)) = w_1 + w_2. \quad (3.33)$$

We can see that weights of evidence are additive and capture the notion of accumulation of evidence.

A simple support function focused on  $A$  with weight  $w$  will be denoted by  $A^w$ . We thus have

$$A^{w_1} \oplus A^{w_2} = A^{w_1+w_2}. \quad (3.34)$$

A mass function is said to be separable if it can be obtained as the combination of simple mass function  $A_1^{w_1}, \dots, A_n^{w_n}$  for some proper non-empty subsets of  $\Omega$ :

$$m = \bigoplus_{i=1}^n A_i^{w_i}, \quad (3.35)$$

We note that this combination is well defined iff

$$\bigcap_{w_i=\infty} A_i \neq \emptyset.$$

A separable mass function generally admits several decompositions as the combination of simple mass functions. As shown by Shafer [77], a particular “canonical” decomposition can be obtained as follows:

$$m = \bigoplus_{A \subseteq \Omega} A^{w(A)}, \quad (3.36)$$

with

$$w(A) = \begin{cases} \sum_{B \subseteq C, B \supseteq A} (-1)^{|B|-|A|} \log Q(B) & \text{if } A \subseteq C, A \neq \emptyset, A \neq C \\ \infty & \text{if } A = C \\ 0 & \text{if } A = \emptyset \text{ or } A \not\subseteq C, \end{cases} \quad (3.37)$$

where  $C$  is the core of  $m$ . Shafer [77] called function  $w : 2^\Omega \rightarrow [0, +\infty]$  defined by (3.37) the *assessment of evidence* associated with  $m$ . As shown in [77], a function  $w : 2^\Omega \rightarrow [0, +\infty]$  is an assessment of evidence (for some separable mass function  $m$ ) if and only if  $w(\emptyset) = 0$ ,  $w$  assigns the value  $+\infty$  to exactly one subset  $C$  of  $\Omega$ , and  $w$  assigns the value zero to every subset of  $\Omega$  not contained in  $C$ . Assessments of evidence defined in that way are in one-to-one correspondence with separable mass functions. A mass function  $m$  is separable if and only if function  $w$  computed using (3.37) is an assessment of evidence. If  $C = \Omega$ , then (3.37) has the following simpler expression,

$$w(A) = \begin{cases} \sum_{B \supseteq A} (-1)^{|B|-|A|} \log Q(B) & \text{if } A \neq \emptyset, A \neq \Omega \\ \infty & \text{if } A = \Omega \\ 0 & \text{if } A = \emptyset. \end{cases} \quad (3.38)$$

We can remark the similarity between (3.38) and (3.8a):  $w$  can be obtained from  $\log Q$  using the same formula that computes  $m$  from  $Q$ .

Not all mass functions are separable. For instance, it is clear that Bayesian mass functions cannot be obtained by combining simple mass functions using Dempster’s rule. However, the class of separable mass function does encompass most mass functions used in practice. Two important cases are considered in the following two propositions [24].

**Proposition 3.1** *Let  $m$  be a mass function with focal sets  $A_1, \dots, A_r$  and  $B$ , such that  $A_k \cap A_\ell = \emptyset$  for all  $k \neq \ell$  and  $\bigcup_{k=1}^r A_k \subseteq B$ . Then  $m$  is*

separable and its assessment of evidence is

$$w(A_k) = -\log\left(\frac{m(B)}{m(A_k) + m(B)}\right), \quad k = 1, \dots, r \quad (3.39a)$$

$$w(B) = +\infty \quad (3.39b)$$

$$w(A) = 0, \quad \forall A \notin \{A_1, \dots, A_r, B\}. \quad (3.39c)$$

*Proof.* We need to show that  $m = A_1^{w(A_1)} \oplus \dots \oplus A_r^{w(A_r)} \oplus B^{w(B)}$ . Let  $m_k = A_k^{w(A_k)}$ , we have

$$m_k(A_k) = \frac{m(A_k)}{m(A_k) + m(B)} \quad (3.40a)$$

$$m_k(\Omega) = \frac{m(B)}{m(A_k) + m(B)}. \quad (3.40b)$$

Let  $m_0 = A_1^{w(A_1)} \oplus \dots \oplus A_r^{w(A_r)}$ . It is easy to see that  $m_0(A_k) = m(A_k)$  for  $k = 1, \dots, r$  and  $m_0(\Omega) = m(B)$ . Combining  $m_0$  with the logical mass function  $m_B$  yields  $m$ .  $\square$

**Proposition 3.2** *Let  $m$  be a consonant mass function, with associated contour function  $\pi_k = pl(\{\omega_k\})$ ,  $k = 1, \dots, n$ . We assume that the elements of  $\Omega = \{\omega_1, \dots, \omega_n\}$  have been arranged in decreasing order of plausibility, i.e., we have*

$$1 = \pi_1 \geq \pi_2 \geq \dots \geq \pi_r > 0$$

and  $\pi_k = 0$  for  $k > r$ . Let  $A_k = \{\omega_1, \dots, \omega_k\}$ ,  $k = 1, \dots, n$ . Then  $m$  is separable and its assessment of evidence is

$$w(A_k) = \log\left(\frac{\pi_k}{\pi_{k+1}}\right), \quad k = 1, \dots, r-1 \quad (3.41a)$$

$$w(A_r) = +\infty \quad (3.41b)$$

$$w(A) = 0, \quad \forall A \notin \{A_1, \dots, A_r\}. \quad (3.41c)$$

*Proof.* Let  $m_k = A_k^{w(A_k)}$  for  $k = 1, \dots, r-1$ . We have

$$m_k(A_k) = 1 - \frac{\pi_{k+1}}{\pi_k}, \quad (3.42a)$$

$$m_k(\Omega) = \frac{\pi_{k+1}}{\pi_k}. \quad (3.42b)$$

After combining the  $r$  mass function  $m_k$ , the mass assigned to  $A_k$  is

$$m(A_k) = m_k(A_k) \prod_{i=1}^{k-1} m_i(\Omega) \quad (3.43a)$$

$$= \left(1 - \frac{\pi_{k+1}}{\pi_k}\right) \prod_{i=1}^{k-1} \frac{\pi_{i+1}}{\pi_i} \quad (3.43b)$$

$$= \frac{\pi_k - \pi_{k+1}}{\pi_k} \pi_k \quad (3.43c)$$

$$= \pi_k - \pi_{k+1} \quad (3.43d)$$

for  $k = 1, \dots, r-1$  and

$$m(A_r) = \prod_{i=1}^{r-1} m_i(\Omega) = \pi_r. \quad (3.44)$$

From Theorem 2.5, this is the consonant mass function with contour function  $\pi_k = pl(\{\omega_k\})$ ,  $k = 1, \dots, n$ .  $\square$

## Exercices

- Let  $m_1$  and  $m_2$  be two mass functions on  $\Omega = \{a, b, c, d\}$  defined as follows

$$m_1(\{a\}) = 0.3 \quad m_1(\{a, c\}) = 0.5 \quad m_1(\{b, c, d\}) = 0.2$$

and

$$m_2(\{b, c\}) = 0.4 \quad m_2(\{a, c, d\}) = 0.5 \quad m_2(\{d\}) = 0.1.$$

Compute the combined mass functions using different combination operators.

- Let  $\Omega = \{a, b\}$ , and let  $m$  and  $m'$  be the following mass functions on  $\Omega$ ,

$$m = \{a\}^\alpha \oplus \{b\}^\beta, \quad m' = \{a\}^{\alpha'} \oplus \{b\}^{\beta'}.$$

- Compute  $m$  and  $m'$ .
  - Compute  $m \oplus m'$ .
- Let  $m$  be a mass function on  $\Omega$  and  $B$  a non-empty subset of  $\Omega$ .

- (a) Express the conditional belief function  $Bel(\cdot|B)$  as a function of  $Bel$ .
  - (b) What does this formula become when  $Bel$  is a probability measure?
4. Let  $m_1$  and  $m_2$  be two consonant mass functions, and let  $Pl_1$  and  $Pl_2$  be the corresponding plausibility measures.
- (a) Show that  $Pl_1 \vee Pl_2 = \max(Pl_1, Pl_2)$  is a plausibility measure.
  - (b) What are the properties of this operator?
  - (c) Using a counterexample, show that  $Pl_1 \vee Pl_2$  may not be a plausibility measure when  $m_1$  and  $m_2$  are not consonant.
5. Prove that the equalities (3.28) hold.



## Chapter 4

# Uncertainty Principles

In many situations, a belief function is only partially specified by some constraints. For instance, assume that we wish to represent an expert's opinion using a belief function. Unless the frame of discernment is very small, it will usually not be practical or even feasible to elicit  $2^{|\Omega|}$  masses or degrees of belief, while maintaining the consistency of the evaluations. Instead, one might elicit some specific aspects of the belief function, such as the contour functions or the plausibility of some propositions. These aspects can then be considered as constraints that should be verified by any belief function representing the expert's opinion. To avoid unwittingly introducing additional information, one should select the *minimally informative* (or *least committed*) belief function satisfying the constraints.

This reasoning mechanism, in which the conclusions are not entailed by the given premises, is sometimes referred to as *ampliative* reasoning [54]. To implement it, one usually applies a *principle of maximal uncertainty*, or *least commitment*. In the Bayesian theory, uncertainty is measured by the Shannon entropy, and the maximum entropy principle is widely used in the probabilistic modeling of information [47]. To apply the maximal uncertainty principle in the belief function framework, we need a way to compare belief functions with respect to their information content. This can be done either qualitatively by defining inclusion relations between belief functions, or quantitatively using uncertainty measures. These approaches will be introduced in Sections 4.1 and 4.2, respectively. Some applications of the Least Commitment Principle (LCP) will then be presented in Section 4.3.

## 4.1 Inclusion relations

Let  $m_1$  and  $m_2$  be two mass functions on the same frame  $\Omega$ . We wish to capture the idea that  $m_1$  is consistent with, but more committed (precise, informative) than  $m_2$ . As we shall see, this can be done in different ways.

In the special case where  $m_1 = m_A$  and  $m_2 = m_B$  are logical mass functions focussed, respectively, on subsets  $A$  and  $B$ , the above condition translates easily to  $A \subseteq B$ . We thus wish to extend the inclusion relation to any belief functions.

### 4.1.1 Belief and commonality-based inclusion relations

In the special case of two logical mass functions mentioned above, we have

$$Bel_A(C) = \begin{cases} 1 & \text{if } A \subseteq C \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

and

$$Bel_B(C) = \begin{cases} 1 & \text{if } B \subseteq C \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

If  $A \subseteq B$ , then  $B \subseteq C \Rightarrow A \subseteq C$ ; consequently,  $Bel_A \geq Bel_B$ .

More generally,  $m_1$  can be considered to be more committed than  $m_2$  if it does not assign a smaller degree of belief to any proposition [29]. This relation will be denoted by  $\sqsubseteq_{Bel}$ :

$$m_1 \sqsubseteq_{Bel} m_2 \text{ iff } Bel_1 \geq Bel_2. \quad (4.3)$$

As  $Pl(A) = 1 - Bel(\bar{A})$  for all  $A$ , this translates to the equivalent condition  $Pl_1 \leq Pl_2$ . We may observe that this relation has a natural interpretation in terms of inclusion of the credal sets (see Section 2.3.3), namely,

$$m_1 \sqsubseteq_{Bel} m_2 \text{ iff } \mathcal{P}(Bel_1) \subseteq \mathcal{P}(Bel_2), \quad (4.4)$$

where  $\mathcal{P}(Bel_i)$  is the credal set of  $Bel_i$ .

Relation  $\sqsubseteq_{Bel}$  is a partial order (i.e., a reflexive, antisymmetric and transitive relation) in the set of mass functions on  $\Omega$ . Its greatest element is the vacuous mass function  $m_?$ , which assigns null degrees of belief to all proper subsets of  $\Omega$ . There is no least element (i.e., no mass function is more committed than any other), but the Bayesian mass functions are minimal elements (for any Bayesian mass function  $m$ , there is no other mass function  $m'$  such that  $m' \sqsubseteq_{Bel} m$ ).

Combining two mass functions  $m_1$  and  $m_2$  using the disjunctive rule (3.26) produces a new mass function  $m_1 \cup m_2$  that is less committed, according to the  $\sqsubseteq_{Bel}$  ordering, than both  $m_1$  and  $m_2$ : we have  $m_1 \sqsubseteq_{Bel} m_1 \cup m_2$  and  $m_2 \sqsubseteq_{Bel} m_1 \cup m_2$ , as a direct consequence of (3.28).

In the special case where  $m_1$  and  $m_2$  are consonant, the condition of inclusion with respect to beliefs can be checked by comparing the contour functions: it can easily be checked that

$$m_1 \sqsubseteq_{Bel} m_2 \Leftrightarrow pl_1 \leq pl_2. \quad (4.5)$$

Instead of reasoning with beliefs, we might as well have reasoned dually with commonalities. Getting back to the special case of two mass logical functions  $m_A$  and  $m_B$ , we have

$$Q_A(C) = \begin{cases} 1 & \text{if } C \subseteq A \\ 0 & \text{otherwise} \end{cases} \quad (4.6)$$

and

$$Q_B(C) = \begin{cases} 1 & \text{if } C \subseteq B \\ 0 & \text{otherwise.} \end{cases} \quad (4.7)$$

If  $A \subseteq B$ , then  $C \subseteq A \Rightarrow C \subseteq B$ ; consequently,  $Q_A \leq Q_B$ .

This observation leads us to define another inclusion relation:  $m_1 \sqsubseteq_Q$  iff  $Q_1 \leq Q_2$ , introduced in [29]. This partial ordering relation has the same properties as  $\sqsubseteq_{Bel}$ , i.e.,  $m_?$  is the greatest element, Bayesian mass functions are minimal elements, and for any two consonant mass functions  $m_1$  and  $m_2$ ,

$$m_1 \sqsubseteq_Q m_2 \Leftrightarrow pl_1 \leq pl_2. \quad (4.8)$$

An immediate consequence is that, given two consonant mass functions  $m_1$  and  $m_2$ ,  $m_1 \sqsubseteq_Q m_2$  iff  $m_1 \sqsubseteq_{Bel} m_2$ . However, this equivalence does not hold in the general case, as shown by the following example.

**Example 4.1** Let  $\Omega = \{a, b, c\}$  and let us consider the following mass functions:

$$\begin{aligned} m_1(\{a\}) &= 1 - \alpha \\ m_1(\{a, b\}) &= 2\alpha - 1 \\ m_1(\Omega) &= 1 - \alpha \end{aligned}$$

and

$$\begin{aligned} m_2(\{a, b\}) &= \alpha \\ m_2(\{a, c\}) &= 1 - \alpha \end{aligned}$$

for some  $\alpha \in (0.5, 1]$ . It can easily be verified that  $m_1 \sqsubseteq_{Bel} m_2$ , whereas  $m_2 \not\sqsubseteq_Q m_1$ .

For any two mass functions  $m_1$  and  $m_2$  with zero degree of conflict, we have  $m_1 \oplus m_2 \sqsubseteq_Q m_1$  and  $m_1 \oplus m_2 \sqsubseteq_Q m_2$ , as a consequence of Theorem 3.1. This observation is consistent with the interpretation of Dempster's rule as a mechanism for pooling evidence:  $m_1 \oplus m_2$  aggregates the two pieces of evidence and, consequently, is more informative than both  $m_1$  and  $m_2$  (assuming there is no conflict).

### 4.1.2 Strong inclusion

Yet another extension of the inclusion relation from subsets to belief functions is the *strong inclusion*, or *specialization* relation [111, 53, 29]. We say that  $m_1$  is strongly included in (is a specialization of)  $m_2$  (which will be denoted by  $m_1 \sqsubseteq_s m_2$ ) iff  $m_1$  can be obtained from  $m_2$  by distributing each mass  $m_2(B)$  for some  $B \subseteq \Omega$  to non-empty subsets of  $B$ :

$$m_1(A) = \sum_{B \supseteq A} S(A, B) m_2(B), \quad (4.9)$$

for all  $A \subseteq \Omega$ , where  $S(A, B)$  is the proportion of  $m_2(B)$  transferred to  $A \subseteq B$ . The numbers  $S(A, B)$  thus verify the following equations:

$$\sum_{A: A \subseteq B} S(A, B) = 1, \quad (4.10)$$

for all  $B \subseteq \Omega$ . We can say that each mass  $m_2(B)$  “flows down” to subsets of  $B$ . Using the vector representation introduced in Section 2.2.3, the terms  $S(A, B)$  can be arranged in an upper triangular square matrix of size  $2^{|\Omega|}$ , called a *specialization matrix* [91].

If  $m_1$  is a specialization of  $m_2$ , then  $m_2$  can be recovered from  $m_1$  by distributing each mass  $m_1(A)$  to supersets of  $A$ , i.e., we have

$$m_2(B) = \sum_{A \subseteq B} G(B, A) m_1(A), \quad (4.11)$$

for all  $A \subseteq \Omega$ , where  $G(B, A)$  is the proportion of  $m_1(A)$  transferred to  $B \supseteq A$ . We say that  $m_2$  is a *generalization* of  $m_1$  and the square matrix with general term  $G(B, A)$  is a *generalization matrix*.

**Example 4.2** Let  $m_1$  and  $m_2$  be the following mass functions on  $\Omega = \{a, b, c\}$ ,

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

We have

$$m_1(\{a\}) = \frac{1}{3}m_2(\{a, b, c\}) = 0.3/3 = 0.1 \quad (4.12a)$$

$$m_1(\{b\}) = m_2(\{b\}) + \frac{1}{3}m_2(\{a, b, c\}) = 0.2 + 0.3/3 = 0.3 \quad (4.12b)$$

$$m_1(\{a, b\}) = m_2(\{a, b\}) + \frac{1}{3}m_2(\{a, b, c\}) = 0.2 + 0.3/3 = 0.3 \quad (4.12c)$$

$$m_1(\{c\}) = \frac{2}{3}m_2(\{a, c\}) = 2 \times 0.3/3 = 0.2 \quad (4.12d)$$

$$m_1(\{a, c\}) = \frac{1}{3}m_2(\{a, c\}) = 0.3/3 = 0.1. \quad (4.12e)$$

The specialization matrix is thus

$$S = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1/3 \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1/3 \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1/3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 2/3 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1/3 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \quad (4.13)$$

where the rows and columns are arranged in the binary order (see Section 2.2.3).

Conversely,  $m_2$  can be obtained from  $m_1$  as follows:

$$m_2(\{b\}) = \frac{2}{3}m_1(\{b\}) = 2 \times 0.3/3 = 0.2 \quad (4.14a)$$

$$m_2(\{a, b\}) = \frac{2}{3}m_1(\{a, b\}) = 2 \times 0.3/3 = 0.2 \quad (4.14b)$$

$$m_2(\{a, c\}) = m_1(\{c\}) + m_1(\{a, c\}) = 0.2 + 0.1 = 0.3 \quad (4.14c)$$

$$\begin{aligned} m_2(\{a, b, c\}) &= m_1(\{a\}) + \frac{1}{3}m_1(\{b\}) + \frac{1}{3}m_1(\{a, b\}) \\ &= 0.1 + 0.3/3 + 0.3/3 = 0.3. \end{aligned} \quad (4.14d)$$

The corresponding generalization matrix is

$$G = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 2/3 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 2/3 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & 1/3 & 1/3 & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}. \quad (4.15)$$

We observe that  $G$  is lower-triangular.

As the two previous ordering relations,  $\sqsubseteq_s$  admits the vacuous mass function  $m_?$  as the unique greatest element and Bayesian mass functions as minimal elements. Furthermore, the following implications hold:

$$m_1 \sqsubseteq_s m_2 \Rightarrow \begin{cases} m_1 \sqsubseteq_{Bel} m_2 \\ m_1 \sqsubseteq_Q m_2. \end{cases} \quad (4.16)$$

These implications are obvious because, as masses flow down to smaller subsets when building  $m_1$  from  $m_2$ , degrees of belief can only increase and commonalities can only decrease. However, the converse implications do not hold: we may have  $m_1 \sqsubseteq_{Bel} m_2$  or  $m_1 \sqsubseteq_Q m_2$  without having  $m_1 \sqsubseteq_s m_2$ . For instance, for the two mass functions of Example 4.1, we can check that  $m_1 \not\sqsubseteq_s m_2$  and  $m_2 \not\sqsubseteq_s m_1$ .

### 4.1.3 Weight-based inclusion

Let  $m_1$  and  $m_2$  be two separable mass functions. We further assume that  $m_1$  and  $m_2$  are *non-dogmatic*, i.e., we assume that  $m_1(\Omega) > 0$  and  $m_2(\Omega) > 0$ . Consequently, their core is  $\Omega$  and, from (3.38), their assessments of evidence  $w_1$  and  $w_2$  verify  $w_i(\Omega) = +\infty$  and  $0 \leq w_i(A) < +\infty$  for all  $A \subset \Omega$  and  $i = 1, 2$ . Keeping in my the interpretation of  $w(A)$  as a weight of evidence pointing to  $A$ , it makes sense to consider that  $m_1$  is more committed than  $m_2$  if  $w_1 \geq w_2$ . We then write  $m_1 \sqsubseteq_w m_2$ . This ordering was introduced in [24], where it was even extended to non-separable mass function, using the general canonical decomposition introduced by Smets in [90]. However, we will only consider the case of separable mass functions here, to keep the exposition simple.

As explained in Section 3.4, assignments of evidence combine additively with Dempster's rule, i.e., the assignment of evidence corresponding to  $m_1 \oplus$

$m_2$  is  $w_1 + w_2$ . It follows that  $m_1 \oplus m_2 \sqsubseteq_w m_1$  and  $m_1 \oplus m_2 \sqsubseteq_w m_2$ . Furthermore, if  $m_1 \sqsubseteq_w m_2$ , there exists a separable function  $m$  such that  $m_1 = m_2 \oplus m$ . This property makes it clear that  $m_1$ , in some sense, aggregates more evidence than  $m_2$ .

## 4.2 Uncertainty measures

Another approach to compare the information content of belief functions is based on the definition of uncertainty measures. As belief functions extend both sets and probability distributions, such measures can be defined as extensions of set-based or probabilistic uncertainty measures. As opposed to the probabilistic case, where there are strong arguments to support the Shannon entropy as the most relevant measure of uncertainty [47], the situation is less clear in belief function theory, in which several notions co-exist. We will review some of the main definitions with their justifications, and show how the minimal commitment principle can be implemented using these measures.

### 4.2.1 Nonspecificity

Assume we receive some piece of information of the form  $\omega \in A$  for some non-empty subset  $A$  of  $\Omega$ . The amount of uncertainty associated with that statement can be measured by the amount of information needed to remove the uncertainty. Such a measure should naturally be a function of the cardinality of  $A$ . Let

$$h : \mathbb{N} \rightarrow \mathbb{R}_+ \tag{4.17}$$

be such a function. Let us consider the following three requirements:

**(H1) Additivity**  $h(r \cdot s) = h(r) + h(s)$ .

**(H2) Monotonicity**  $h(s) < h(s + 1)$ .

**(H3) Normalization**  $h(2) = 1$ .

Requirement *H2* is quite natural and *H3* is a matter of convention. The only non-trivial axiom is *H1*; it has the following meaning. Consider a partition of  $\Omega$  into  $r$  subsets of  $s$  elements. Characterizing an element of  $\Omega$  requires the amount  $h(r \cdot s)$  of information. However, we could also proceed in two steps: first, we could characterize the subset to which the element belongs (requiring an amount  $h(r)$  of information, and then characterize the element

in this subset (with required information  $h(s)$ ). The equivalence of the two methods leads to Axiom *H1*.

It can be shown [54] that the only function  $h$  verifying these three axioms is defined by

$$h(n) = \log_2 n. \quad (4.18)$$

The function  $H : 2^\Omega \setminus \emptyset \rightarrow \mathbb{R}_+$  defined by  $H(A) = \log_2 |A|$  is called the Hartley function. Its range is  $[0, \log_2 |\Omega|]$ . We can observe that  $H(A)$  is equal to the Shannon entropy of the uniform probability distribution on  $A$ .

As a natural extension from sets to mass function, we can compute the average value of the Hartley function for all focal sets,

$$N(m) = \sum_{\emptyset \neq A \subseteq \Omega} m(A) \log_2 |A|, \quad (4.19)$$

which is called the *nonspecificity* of  $m$  [28]. The range of  $N$  is the same as that of  $H$ , and the following implication holds:

$$m_1 \sqsubseteq_s m_2 \Rightarrow N(m_1) \leq N(m_2), \quad (4.20)$$

where  $\sqsubseteq_s$  denotes the strong inclusion relation (see Section 4.1.2).

The nonspecificity was shown by Ramer [72] to be the only function satisfying some rationality axioms. Before introducing these axioms, we need to define the notion of *non-interactive marginal mass functions*. Let  $m^{\Omega \times \Theta}$  be a mass function on a product frame  $\Omega \times \Theta$ . The marginals<sup>1</sup> of  $m^{\Omega \times \Theta}$  on  $\Omega$  and  $\Theta$  are the mass functions obtained by transferring each mass  $m^{\Omega \times \Theta}(C)$  to the projections of  $C$  on  $\Omega$  and  $\Theta$ , respectively. The marginals  $m^\Omega$  and  $m^\Theta$  are said to be non-interactive if, for all  $C \subseteq \Omega \times \Theta$ ,

$$m^{\Omega \times \Theta}(C) = \begin{cases} m^\Omega(A)m^\Theta(B) & \text{if } C = A \times B, \\ 0 & \text{otherwise.} \end{cases} \quad (4.21)$$

Otherwise, they are said to be interactive.

**(N1) Additivity for non-interactive mass functions** Let  $m^{\Omega \times \Theta}$  be a joint mass function with non-interactive marginals  $m^\Omega$  and  $m^\Theta$ , then

$$N(m^{\Omega \times \Theta}) = N(m^\Omega) + N(m^\Theta). \quad (4.22)$$

**(N2) Subadditivity for interactive mass functions** Let  $m^{\Omega \times \Theta}$  be a joint mass function with interactive marginals  $m^\Omega$  and  $m^\Theta$ , then

$$N(m^{\Omega \times \Theta}) \leq N(m^\Omega) + N(m^\Theta). \quad (4.23)$$

---

<sup>1</sup>The notion of marginal mass function will be further studied in Chapter 6.



**(N3) Normalization**  $N(m) = 1$  when  $m(A) = 1$  and  $|A| = 2$ .

**(N4) Symmetry**  $N$  is invariant with respect to permutations of values of the mass functions within each group of subsets of  $\Omega$  that have equal cardinalities.

**(N5) Branching**  $N(m) = N(m_1) + N(m_2)$  for any three bodies of evidence with focal sets

$$\mathcal{F} = \{A, B, C, \dots\}, \mathcal{F}_1 = \{A_1, B, C, \dots\}, \mathcal{F}_2 = \{A, B_1, C_1, \dots\},$$

where  $A_1 \subseteq A$ ,  $B_1 \subseteq B$ ,  $C_1 \subseteq C$ , etc. and  $|A_1| = |B_1| = |C_1| = \dots = 1$ , and

$$m(A) = m_1(A_1) = m_2(A),$$

$$m(B) = m_1(B) = m_2(B_1),$$

$$m(C) = m_1(C) = m_2(C_1), \text{ etc.}$$

Nonspecificity is obviously a well justified measure of the uncertainty of a belief function. However, we can remark that  $N(m) = 0$  for any Bayesian mass function. This observation shows that nonspecificity only measures one aspect of uncertainty, related to imprecision. Another aspect of uncertainty is related to *conflict*, i.e., the fact that evidence points to several totally or partially disjoint focal sets. This aspect may be described using other measures that extend the Shannon entropy.

### 4.2.2 Conflict measures

Seeing a mass function as the density of a random set [67], we could define its entropy as

$$S(m) = - \sum_{A \subseteq \Omega} m(A) \log_2 m(A). \quad (4.24)$$

The problem with that definition is that it does not take into account the structure of the focal sets. For instance, the two mass functions  $m_1(\{\omega_1\}) = 0.5$ ,  $m_1(\{\omega_2\}) = 0.5$  and  $m_2(\{\omega_1\}) = 0.5$ ,  $m_2(\{\omega_1, \omega_2\}) = 0.5$  have the same entropy, yet the latter has obviously less uncertainty (and less conflict) than the former.

Several alternative generalizations of the Shannon entropy to belief functions have been proposed. One of them is *dissonance* [110], defined by

$$E(m) = - \sum_{A \subseteq \Omega} m(A) \log_2 Pl(A) \quad (4.25a)$$

$$= - \sum_{A \subseteq \Omega} m(A) \log_2 \left( 1 - \sum_{B \cap A = \emptyset} m(B) \right). \quad (4.25b)$$

The term

$$K(A) = \sum_{B \cap A = \emptyset} m(B) \quad (4.26)$$

in (4.25b) can be interpreted as measuring the degree to which the evidence conflicts with focal set  $A$ . The term

$$- \log_2[1 - K(A)] \quad (4.27)$$

is strictly increasing with  $K(A)$  and extends its range to  $[0, +\infty)$ . Dissonance can thus be seen as the mean value of the conflict among focal sets of a piece of evidence.

Several variants have been proposed. For instance, it has been proposed to replace  $K(A)$  by the following conflict measure

$$CON(A) = \sum_{\emptyset \neq B \subseteq \Omega} m(B) \frac{|A \setminus B|}{|A|}, \quad (4.28)$$

which takes into account the proportion of elements of  $A$  not included in  $B$ . Replacing  $K(A)$  with  $CON(A)$ , we get a new uncertainty measure called *strife* [54], defined by the formula:

$$ST(m) = - \sum_{A \subseteq \Omega} m(A) \log_2 \left( 1 - \sum_{\emptyset \neq B \subseteq \Omega} m(B) \frac{|A \setminus B|}{|A|} \right) \quad (4.29a)$$

$$= - \sum_{A \subseteq \Omega} m(A) \log_2 \sum_{B \cap A \neq \emptyset} m(B) \frac{|A \cap B|}{|A|}. \quad (4.29b)$$

### 4.2.3 Total uncertainty measures

Some authors have proposed uncertainty measures addressing the two dimensions of uncertainty mentioned in the previous sections, i.e., nonspecificity (or imprecision), and conflict. For example, the total uncertainty in a belief

function may be defined the *aggregate uncertainty* [60, 41], defined as the maximum value of the Shannon entropy over all compatible probabilities:

$$AU(Bel) = \max_{\mathcal{P}(Bel)} \left( - \sum_{\omega \in \Omega} p(\omega) \log_2 p(\omega) \right). \quad (4.30)$$

The range of  $AU$  is  $[0, \log_2 |\Omega|]$ . It boils down to the Shannon entropy when  $Bel$  is Bayesian and to the Hartley function when  $Bel$  is logical. It is also subadditive for interactive marginal belief functions and additive for interactive ones. However, there is no proof that it is the only measure verifying all these properties.

A different approach was proposed by Smets [87], who argued that a measure of the information content of a mass function should be additive with respect to Dempster's rule in the case where there is no conflict. Let  $m_1$  and  $m_2$  be two mass functions, and let  $Q_1$  and  $Q_2$  be the corresponding commonality functions. We know that, when the degree of conflict is null, we have  $Q_1 \oplus Q_2 = Q_1 Q_2$ . The requirement

$$I(m_1 \oplus m_2) = I(m_1) + I(m_2) \quad (4.31)$$

as well as some other trivial requirements lead to

$$I(m) = - \sum_{A \subseteq \Omega} c(A) \log Q(A), \quad (4.32)$$

where  $c(A)$  is a constant depending on  $A$ . Expression (4.32) is well-defined as long as  $m(\Omega) > 0$ , in which case  $Q(A) > 0$  for all  $A \subseteq \Omega$ . Choosing  $c(A) = 1$ , we get

$$I(m) = - \sum_{A \subseteq \Omega} \log Q(A). \quad (4.33)$$

It is obvious that

$$m_1 \sqsubseteq_Q m_2 \Rightarrow I(m_1) \geq I(m_2). \quad (4.34)$$

### 4.3 Applications

Uncertainty measures can be used to implement the minimal commitment principle, especially in situations when the least committed belief function (according to some inclusion relation) satisfying some constraints does not exist or is hard to find. The nonspecificity measure is specially convenient as it is linear in the masses: when the constraints are themselves linear, finding a least committed mass functions satisfying the constraints is a linear programming problem and can be solved very efficiently. We will study several examples.

### 4.3.1 Least committed belief function from consonant sets

Assume that we wish to elicit an expert's opinion on given question defined on a frame  $\Omega$ . We may do it by asking him to give a strictly increasing sequence

$$A_1 \subset A_2 \subset \dots \subset A_n$$

of subsets of  $\Omega$  with their degrees of belief

$$0 < \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n = 1.$$

The least committed mass function (according to any of the ordering relations  $\sqsubseteq_{Bel}$ ,  $\sqsubseteq_Q$  and  $\sqsubseteq_s$ ) can be constructed as follows. To satisfy the constraint  $Bel(A_1) = \alpha_1$ , we need to distribute a mass  $\alpha_1$  to some subsets of  $A_1$ . The largest such subset is  $A_1$  itself. We thus get  $m(A_1) = \alpha_1$ . To satisfy the constraint  $Bel(A_2) = \alpha_2$ , we now need to distribute a mass  $\alpha_2 - \alpha_1$  to some subsets of  $A_2$  that are not subsets of  $A_1$ . The largest such subset is  $A_2$ . We thus get  $m(A_2) = \alpha_2 - \alpha_1$ . By pursuing this line of reasoning, we get the following consonant mass function:

$$\begin{aligned} m(A_1) &= \alpha_1 \\ m(A_2) &= \alpha_2 - \alpha_1 \\ &\vdots \\ m(A_i) &= \alpha_i - \alpha_{i-1} \\ &\vdots \\ m(A_n) &= 1 - \alpha_{n-1}. \end{aligned}$$

### 4.3.2 Conditional embedding

As another example of partial information about a belief function, assume that a source gives us a mass function  $m_0$  representing evidence about some question  $Q$  defined on frame  $\Omega$ , assuming that some proposition  $A \subset \Omega$  holds. This mass function can be interpreted as a conditional mass function  $m(\cdot|A)$  obtained by conditioning some unknown mass function  $m$  by  $A$  (see Section 3.2.3). However, there usually exist several mass functions  $m$  verifying this property. Let  $\mathcal{M}_A(m_0)$  be the set of mass functions such that  $m(\cdot|A) = m_0$ . The least committed element in  $\mathcal{M}_A(m_0)$  (according to the three previous inclusion relations) can be found as follows.

Let  $m$  be a mass function in  $\mathcal{M}_A(m_0)$ . Any mass  $m_0(B)$  for some  $B \subseteq A$  is obtained from  $m$  by transferring masses  $m(B \cup C)$  to  $B$ , for some  $C \subseteq \bar{A}$ :

$$m_0(B) = \sum_{C \subseteq \bar{A}} m(B \cup C). \quad (4.35)$$

The largest possible  $C$  is  $\bar{A}$  itself. The least committed mass function  $m_{\text{LC}}$  in  $\mathcal{M}_A(m_0)$  is thus obtained by transferring each mass  $m_0(B)$  to  $B \cup \bar{A}$ :

$$m_{\text{LC}}(D) = \begin{cases} m_0(B) & \text{if } D = B \cup \bar{A}, \\ 0 & \text{otherwise.} \end{cases} \quad (4.36)$$

The operation that maps  $m_0$  to  $m_{\text{LC}}$  is called *deconditioning*, or *conditional embedding*.

### 4.3.3 Partial belief specifications

Consider a problem similar to that presented in Section 4.3.1, where we are given the degrees of belief for  $r$  subsets  $A_1, \dots, A_r$ . However, we no longer impose any structure on these sets. Let  $\alpha_i = \text{Bel}(A_i)$ ,  $i = 1, \dots, r$ . The least specific mass function verifying these constraints can be found by solving the following linear program:

$$\max_m \sum_{A \subseteq \Omega} m(A) \log_2 |A| \quad (4.37a)$$

under the constraints:

$$\sum_{B \subseteq A_i} m(B) = \alpha_i, \quad i = 1, \dots, r \quad (4.37b)$$

$$\sum_{A \subseteq \Omega} m(A) = 1 \quad (4.37c)$$

$$m(\emptyset) = 0. \quad (4.37d)$$

### 4.3.4 Combination of mass functions with unknown dependence

Let us consider two sources  $(S_1, \mu_1, \Gamma_1)$  and  $(S_2, \mu_2, \Gamma_2)$  generating mass functions  $m_1$  and  $m_2$ . When taking into account both items of evidence jointly, we must consider a joint probability measure  $\mu_{12}$  on  $S_1 \times S_2$ , whose marginals are  $\mu_1$  and  $\mu_2$ .

Let  $A_1, \dots, A_r$  denote the focal sets of  $m_1$ ,  $B_1, \dots, B_s$  the focal sets of  $m_2$ ,  $p_i = m_1(A_i)$ ,  $i = 1, \dots, r$ ,  $q_j = m_2(B_j)$ ,  $j = 1, \dots, s$ , and

$$p_{ij} = \mu_{12}(\{(s_1, s_2) \in S_1 \times S_2 | \Gamma_1(s_1) = A_i, \Gamma_2(s_2) = B_j\}). \quad (4.38)$$

Assuming both sources to be reliable, and that the degree of conflict

$$\kappa = \sum_{A_i \cap B_j = \emptyset} p_{ij} \quad (4.39)$$

is not equal to 1, the combined mass function  $m$  has the following expression

$$m(A) = \sum_{A_i \cap B_j = A} p_{ij}^*, \quad (4.40)$$

for all  $A \subseteq \Omega$ ,  $A \neq \emptyset$ , with  $p_{ij}^* = p_{ij}/(1 - \kappa)$ .

When the dependence between the two sources is unknown, the  $p_{ij}$ 's are unknown, but we can find the least specific mass function of the form (4.40) by solving the following non-linear optimization problem:

$$\max_{p_{ij}} \sum_{\{(i,j) | A_i \cap B_j \neq \emptyset\}} p_{ij}^* \log_2 |A_i \cap B_j| \quad (4.41a)$$

under the constraints:

$$\sum_{i,j} p_{ij} = 1 \quad (4.41b)$$

$$\sum_i p_{ij} = q_j, \quad j = 1, \dots, s \quad (4.41c)$$

$$\sum_j p_{ij} = p_i, \quad i = 1, \dots, r. \quad (4.41d)$$

We note that, when there is no pair  $(A_i, B_j)$  of focal sets such that  $A_i \cap B_j = \emptyset$ , we have  $p_{ij}^* = p_{ij}$  for all  $i$  and  $j$ , and the above optimization problem becomes linear.

## Exercises

1. Compare the mass functions of Example 4.1 using the relations  $\sqsubseteq_{Bel}$ ,  $\sqsubseteq_Q$  and  $\sqsubseteq_s$ .
2. Consider the following mass functions on a frame  $\Omega = \{a, b, c\}$ :

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

Is  $m_1$  a specialization or a generalization of  $m_2$ ? Depending on the answer to the previous question, compute the corresponding specialization and generalization matrices.

3. Let  $\Omega = \{a, b, c\}$  be a frame of discernment. Assume that an expert gives us the following information:

$$Bel(\{a, b\}) = 0.4,$$

$$Pl(\{b, c\}) = 0.8,$$

$$Pl(\{a, c\}) = 0.7.$$

Show that a mass function  $m$  with maximum nonspecificity  $N(m)$  and verifying the above equalities can be found by solving the following linear optimization problem,

$$\max \mathbf{w}'\mathbf{m}$$

$$\text{such that } \mathbf{A}\mathbf{m} = \mathbf{b},$$

for some matrix  $\mathbf{A}$  and some vectors  $\mathbf{w}$  and  $\mathbf{b}$ .





## Chapter 5

# The cautious rule of combination



## Chapter 6

# Varying the granularity of the frame

As already mentioned in Section 2.1.1, the definition of a frame of discernment is, to a large extent, a matter of convention, as its granularity is often a matter of choice. Different sources of information may provide evidence represented in frames of different granularities. Consider, for example, a multi-sensor system for road scene understanding such as described in [109]. Some sensor or image processing algorithm may detect if an object is a pedestrian or not, while some others may provide more detailed information about the nature of the object (such as two or four-wheel vehicle, etc.). Also, when several variables are defined, we may receive evidence about different subsets of variables, and we may wish to express the result of the analysis according to some specific subset containing the variables of interest [85]. In evidential reasoning with uncertain information, we thus have to express belief functions in different frames with varying granularities.

### 6.1 Refinement and coarsening

Let  $\Omega$  and  $\Theta$  be two frames of discernment. We say that  $\Omega$  is a *refinement* of  $\Theta$  (or, equivalently,  $\Theta$  is a *coarsening* of  $\Omega$ ) if the elements of  $\Omega$  can be obtained by splitting some or all of the elements of  $\Theta$  (Figure 6.1) [77]. Formally,  $\Omega$  is a refinement of a frame  $\Theta$  iff there is a mapping  $\rho : 2^\Theta \rightarrow 2^\Omega$  (called a *refining*) such that:

- $\{\rho(\{\theta\}), \theta \in \Theta\} \subseteq 2^\Omega$  is a partition of  $\Omega$ , and
- For all  $A \subseteq \Omega$ ,  $\rho(A) = \bigcup_{\theta \in A} \rho(\{\theta\})$ .

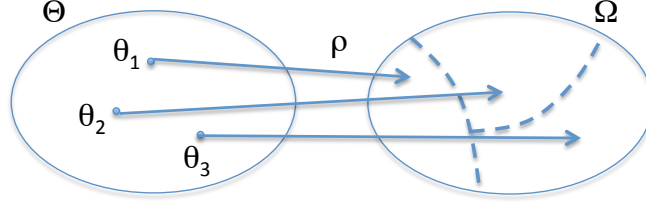


Figure 6.1: Refinement of a frame of discernment.

Let  $m^\Theta$  be a mass function representing some piece of evidence expressed in the frame  $\Theta$ , and let  $\Omega$  be a refinement of  $\Theta$ . We can carry  $m^\Theta$  from  $\Theta$  to  $\Omega$  by transferring each mass  $m^\Theta(A)$  to  $\rho(A)$ . The resulting mass function is denoted by  $m^{\Theta\uparrow\Omega}$  and is called the *vacuous extension* of  $m^\Theta$  in  $\Omega$ : for all  $B \subseteq \Omega$ ,

$$m^{\Theta\uparrow\Omega}(B) = \begin{cases} m^\Theta(A) & \text{if } B = \rho(A), A \subseteq \Theta, \\ 0 & \text{otherwise.} \end{cases} \quad (6.1)$$

Conversely, given a mass function  $m^\Omega$  on  $\Omega$ , how to express it in the coarser frame  $\Theta$ ? Here, the solution is not so obvious because the mapping  $\rho$  is not invertible. However, we can define two generalized inverses of  $\rho$ . Let  $\underline{\rho}^{-1}$  and  $\bar{\rho}^{-1}$  be two mappings from  $2^\Omega$  to  $2^\Theta$  defined as follows:

$$\underline{\rho}^{-1}(B) = \{\theta \in \Theta \mid \rho(\{\theta\}) \subseteq B\}, \quad (6.2a)$$

$$\bar{\rho}^{-1}(B) = \{\theta \in \Theta \mid \rho(\{\theta\}) \cap B \neq \emptyset\}, \quad (6.2b)$$

for any subset  $B$  of  $\Omega$ . The subsets  $\underline{\rho}^{-1}(B)$  and  $\bar{\rho}^{-1}(B)$  are called, respectively, the *inner reduction* and the *outer reduction* of  $B$  [77]. When computing the image by  $\rho$  of the inner or the outer reduction of  $B$ , we do not recover  $B$  in general. The following relations hold:

$$\rho[\underline{\rho}^{-1}(B)] \subseteq B \subseteq \rho[\bar{\rho}^{-1}(B)], \quad (6.3)$$

and the inclusions may be strict.

Let us now assume that we have a mass function  $m^\Omega$  defined in  $\Omega$ . In principle, it can be carried to  $\Theta$  in two ways, i.e., by transferring each mass  $m^\Omega(B)$  to the inner reduction or to the outer reduction of  $B$ . In the latter case, the resulting mass function is denoted by  $m^{\Omega\downarrow\Theta}$  and is called the *restriction* of  $m^\Omega$  in  $\Theta$ : for all subset  $A$  of  $\Theta$ ,

$$m^{\Omega\downarrow\Theta}(A) = \sum_{\bar{\rho}^{-1}(B)=A} m^\Omega(B). \quad (6.4)$$

We may observe that, in the process of carrying  $m^\Omega$  from  $\Omega$  to the coarser frame  $\Theta$ , some information may be lost. In particular, if  $m^{\Omega \downarrow \Theta}$  is carried back to  $\Omega$ , we will not recover  $m^\Omega$  in general. The resulting mass function will usually be strictly less informative than  $m^\Omega$  according to the strong inclusion relation  $\sqsubseteq_s$  (cf. Section 4.1.2), because any mass  $m^\Omega(B)$  initially assigned to  $B$  is now assigned to a superset  $\rho[\bar{\rho}^{-1}(B)]$ . This actually makes sense: since some information is lost, we get a less informative mass function. This is the reason why the outer reduction  $\bar{\rho}^{-1}$  is used in (6.4).

## 6.2 Special case of product spaces

### 6.2.1 Marginalization and vacuous extension

Let us now assume that we have two frames  $\Omega_X$  and  $\Omega_Y$  related to two different questions about, e.g., the values of two unknown variables  $X$  and  $Y$ . Let  $\Omega_{XY} = \Omega_X \times \Omega_Y$  be the product space. It is a refinement of both  $\Omega_X$  and  $\Omega_Y$ . For instance, we can define the following refining  $\rho$  from  $2^{\Omega_X}$  to  $2^{\Omega_X \times \Omega_Y}$ :

$$\rho(A) = A \times \Omega_Y, \quad (6.5)$$

for all  $A \subseteq \Omega_X$ . The set  $\rho(A)$  is called the *cylindrical extension* of  $A$  in  $\Omega_{XY}$  and is denoted by  $A \uparrow \Omega_{XY}$  (see Section 1.2.1).

The vacuous extension of a mass function  $m^X$  from  $\Omega_X$  to  $\Omega_{XY}$  is obtained by transferring each mass  $m^{\Omega_X}(B)$  for any subset  $B$  of  $\Omega_X$  to the cylindrical extension of  $B$ :

$$m^{X \uparrow \Omega_{XY}}(A) = \begin{cases} m^X(B) & \text{if } A = B \times \Omega_Y \\ 0 & \text{otherwise.} \end{cases} \quad (6.6)$$

Conversely, let  $m^{XY}$  be a joint mass function on the product space  $\Omega_{XY}$ . Typically, such a mass function represents partial knowledge about the relation between variables  $X$  and  $Y$ . Now, assume that we are only interested in evidence about  $\Omega_X$ . We then have to compute the restriction of  $m^{XY}$  to the coarser frame  $\Omega_X$ . As explained in Section 6.1, it is obtained by transferring each mass  $m^{XY}(B)$  to the outer reduction of  $B$ , defined as

$$\bar{\rho}^{-1}(B) = \{x \in \Omega_X \mid \rho(\{x\}) \cap B \neq \emptyset\} \quad (6.7a)$$

$$= \{x \in \Omega_X \mid \exists y \in \Omega_Y, (x, y) \in B\} \quad (6.7b)$$

$$= B \downarrow \Omega_X, \quad (6.7c)$$

where  $B \downarrow \Omega_X$  denotes the *projection* of  $B$  on  $\Omega_X$  (see Section 1.2.1). We thus have

$$m^{XY \downarrow X}(A) = \sum_{B \downarrow \Omega_X = A} m^{XY}(B), \quad (6.8)$$

for all  $A \subseteq \Omega_X$ . The mass functions  $m^{XY \downarrow X}$  and  $m^{XY \downarrow Y}$  are called the *marginals* of  $m^{XY}$  and the operation that computes the marginals from a joint mass function is called *marginalization*.

We can observe that this operation extends both set projection and marginalization of joint probability distributions:

- If  $m_B^{XY}$  is a logical mass function with focal set  $B \subseteq \Omega_{XY}$ , its marginal  $m_B^{XY \downarrow X}$  is still a logical mass function with focal set  $B \downarrow \Omega_X$ .
- If  $m^{XY}$  is a Bayesian mass function, then  $m^{XY \downarrow X}$  is still Bayesian and it is defined by

$$m^{XY \downarrow X}(\{x\}) = \sum_{y \in \Omega_Y} m^{XY}(\{x, y\}), \quad (6.9)$$

which corresponds to probabilistic marginalization.

## 6.2.2 Application to evidential reasoning

Most problems in engineering or economics can be modeled by defining variables and relations between variables. Based on partial information about some variables, the problem is then to infer the values of variables of interest. This problem can be cast in the belief function framework, as relations are sets and can thus be represented by joint mass functions. The three fundamental operations in evidential reasoning are Dempster's rule, marginalization and vacuous extension. For instance, assume for simplicity that we have only two variables  $X$  and  $Y$  and we have:

- Partial knowledge of  $X$  formalized as a mass function  $m^X$ , and
- A joint mass function  $m^{XY}$  representing an *uncertain relation* between  $X$  and  $Y$ .

These two pieces of evidence can be combined by vacuously extending  $m^X$  to  $\Omega_{XY}$  and combining  $m^{X \uparrow XY}$  with  $m^{XY}$ . The combined joint mass function can then be marginalized on  $\Omega_Y$ . Formally,

$$m^Y = \left( m^{X \uparrow XY} \oplus m^{XY} \right)^{\downarrow Y}. \quad (6.10)$$

To simplify the notation, we may assume implicitly that the mass functions being combined are expressed in the coarsest common refinement. For instance,  $m^X \oplus m^Y$  is understood to be  $m^{X \uparrow XY} \oplus m^{Y \uparrow XY}$ . With this convention, (6.10) can be written more compactly as

$$m^Y = (m^X \oplus m^{XY}) \downarrow^Y. \quad (6.11)$$

Equation (6.11) is clearly a generalization of (1.8).

We can remark that these operations become intractable when the number of variables and/or the size of the frames of discernment become very large, but efficient algorithms exist to carry out the operations in frames of minimal dimensions [83, 4, 86].

In Section 6.4 below, we will describe two important applications of this mechanism. However, before that, we need to come back to the notions of conditioning and deconditioning in the context of multiple frames of discernment.

### 6.3 Conditioning and deconditioning

In Section 3.2.3, we introduced the notion of conditioning as a special case of Dempster's rule. Given a mass function  $m$  on  $\Omega$ , a subset  $B$  of  $\Omega$  and  $m_B$  the logical mass function such as  $m_B(B) = 1$ , the conditional belief function  $m(\cdot|B)$  is:

$$m(\cdot|B) = m \oplus m_B. \quad (6.12)$$

An inverse operation, called deconditioning or conditional embedding, was later introduced in Section 4.3.2 as a consequence of the Least Commitment Principle. Given a mass function  $m$  on  $\Omega$  whose core is included in  $B$ , its deconditioning yields the mass function  $m'$  defined by

$$m'(C) = \begin{cases} m(A) & \text{if } C = A \cup \overline{B} \text{ for some } A \subseteq B, \\ 0 & \text{otherwise,} \end{cases} \quad (6.13)$$

for any  $D \subseteq \Omega$ .

Let us now assume that  $m^{XY}$  is a mass function on the product space  $\Omega_X \times \Omega_Y$  and  $B$  is a subset of  $\Omega_Y$ . The conditional mass function  $m^X(\cdot|B)$  is defined by combining  $m^{XY}$  with  $m_B^{Y \uparrow XY}$  (where  $m_B^Y$  is the logical mass function focussed on  $B$ ), and marginalizing the result on  $X$ :

$$m^X(\cdot|B) = (m^{XY} \oplus m_B^{Y \uparrow XY}) \downarrow^X. \quad (6.14)$$

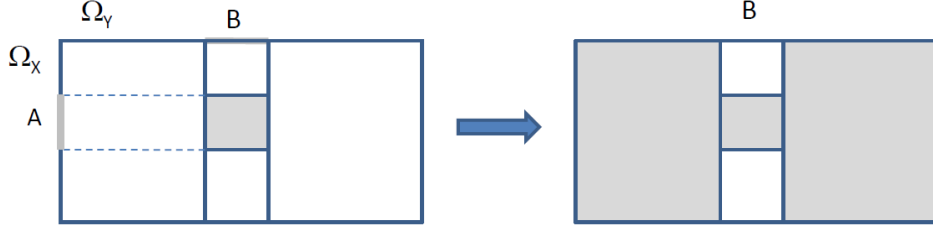


Figure 6.2: Conditional embedding operation. The mass on  $m^X(A|B)$  is transferred to  $(A \times \Omega_Y) \cup (\Omega_X \times \bar{B})$ .

Conversely, let  $m^X(\cdot|B)$  represents your beliefs on  $\Omega_X$  conditionally on  $B$  for some  $B \subseteq \Omega_Y$ , i.e., in a context where  $Y \in B$  holds. There are usually many mass functions on  $\Omega_X \times \Omega_Y$ , whose conditioning on  $B$  yields  $m^X(\cdot|B)$ . The least committed one can be obtained by vacuously extending  $m^X(\cdot|B)$  in  $\Omega_X \times \Omega_Y$  and deconditioning using (6.13). As shown in [89], the results is

$$m^{XY}(C) = \begin{cases} m^X(A|B) & \text{if } C = (A \times \Omega_Y) \cup (\Omega_X \times \bar{B}) \text{ for some } A \subseteq \Omega_X, \\ 0 & \text{otherwise.} \end{cases} \quad (6.15)$$

## 6.4 Applications

### 6.4.1 Discounting

Let us assume that we receive a mass function  $m_0^\Omega$  from a source  $S$ , describing some information about  $\omega$ . However, this information is not fully reliable or not fully relevant because, e.g., it is provided by a possibly faulty sensor, the measurement was performed in unfavorable experimental condition, or the information is related to a situation or an object that only has some similarity with the situation or the object considered. By combining information provided by the source with metaknowledge about the reliability of the source, we get a new, less informative mass function. This operation is called *discounting* [77, 89].

Let  $\mathcal{R} = \{R, NR\}$  be the set of possible states of the source, i.e., reliable or not reliable. Let us assume we have a Bayesian mass function  $m^{\mathcal{R}}$  on  $\mathcal{R}$ ,

$$m^{\mathcal{R}}(\{R\}) = 1 - \alpha \quad (6.16a)$$

$$m^{\mathcal{R}}(\{NR\}) = \alpha, \quad (6.16b)$$



for some  $\alpha \in [0, 1]$ . If  $S$  is reliable, we can adopt  $m_0^\Omega$  to represent our belief,

$$m^\Omega(\cdot|R) = m_0^\Omega. \quad (6.17)$$

If  $S$  is not reliable, the information it provides cannot be taken into account, and our mass function should be vacuous:

$$m^\Omega(\Omega|NR) = 1. \quad (6.18)$$

Therefore, we have two non-vacuous pieces of evidence,  $m^\mathcal{R}$  and  $m^\Omega(\cdot|R)$ . Vacuously extending  $m^\mathcal{R}$  on  $\Omega \times \mathcal{R}$  results in the following mass function:

$$m_1^{\Omega \times \mathcal{R}}(\Omega \times \{R\}) = m^\mathcal{R}(\{R\}) = 1 - \alpha, \quad (6.19a)$$

$$m_1^{\Omega \times \mathcal{R}}(\Omega \times \{NR\}) = m^\mathcal{R}(\{NR\}) = \alpha. \quad (6.19b)$$

Now, deconditioning  $m^\Omega(\cdot|R)$  yields a second mass function on the product space,

$$m_2^{\Omega \times \mathcal{R}}((A \times \mathcal{R}) \cup (\Omega \times \{NR\})) = m_0^\Omega(A), \quad (6.20)$$

for all  $A \subseteq \Omega$ . Combining  $m_1^{\Omega \times \mathcal{R}}$  and  $m_2^{\Omega \times \mathcal{R}}$  and marginalizing the result on  $\Omega$  yields the *discounted* mass function,

$$\alpha m^\Omega(A) = \begin{cases} (1 - \alpha)m_0^\Omega(A) & \text{if } A \neq \Omega, \\ (1 - \alpha)m_0^\Omega(\Omega) + \alpha & \text{if } A = \Omega. \end{cases} \quad (6.21)$$

The coefficient  $\alpha$  is called the *discount rate*. The mass function  $\alpha m^\Omega$  is simply a weighted sum of  $m_0^\Omega$  and the vacuous mass function  $m_?^\Omega$ :

$$\alpha m^\Omega = (1 - \alpha)m_0^\Omega + \alpha m_?^\Omega. \quad (6.22)$$

The corresponding belief function has the following simple expression,

$$\alpha Bel^\Omega(A) = \begin{cases} (1 - \alpha)Bel_0^\Omega(A) & \text{if } A \neq \Omega \\ 1 & \text{otherwise.} \end{cases} \quad (6.23)$$

### 6.4.2 Generalized Bayes' Theorem

The Generalized Bayes' Theorem (GBT) is due to Smets [89]. It is an extension of Bayes' theorem when probabilities are replaced by general belief functions.

Let us assume that we have two variables  $X$  and  $Y$  defined on frames  $\Omega_X$  and  $\Omega_Y$ . Let  $\Omega_Y = \{y_1, \dots, y_n\}$  and  $m^X(\cdot|y_k) = m_k^X$ , for  $k = 1, \dots, n$

be a mass functions representing our belief about  $X$  in a context where  $Y = y_i$  holds. These conditional mass functions are assumed to be provided by independent sources. Based on this conditional knowledge, we want to derive a mass function representing our belief about  $Y$  when  $X \in A$  holds, for some  $A \subseteq \Omega_X$ .

For instance, assume that  $\Omega_X$  is a set of symptoms and  $\Omega_Y$  is set of health states of a patient. Based on evidence from past observations, we can construction a conditional mass function  $m_k^X$  on the set of symptoms, for each of the patient's states  $y_k$ . If the datasets for two states  $y$  and  $y'$  are not overlapping, we can consider the mass functions  $m_k^X$  for  $k = 1, \dots, n$  as independent. The diagnosis problem is to derive a belief function  $m^Y(\cdot|A)$  on the patient's state, knowing that the patient's symptom belongs to a subset  $A$  of  $\Omega_X$ .

This problem can be solved by applying the following steps [89]:

1. Decondition each conditional mass function  $m_k^X$  in  $\Omega_X \times \Omega_Y$ ; the resulting mass function is such that

$$m_k^{XY}(C \times \Omega_Y \cup \Omega_X \times \overline{\{y\}}) = m_k^X(C) \quad (6.24)$$

for all  $C \subseteq \Omega_X$ .

2. Condition each mass function  $m_k^{XY}$  by  $A \times \Omega_Y$  and marginalize on  $\Omega_Y$ . We get the following simple mass function on  $\Omega_Y$ ,

$$m_k^Y(\overline{\{y_k\}}|A) = 1 - Pl_k^X(A) \quad (6.25a)$$

$$m_k^Y(\Omega_Y|A) = Pl_k^X(A), \quad (6.25b)$$

where  $Pl_k^X$  is the plausibility function corresponding to  $m_k^X$ . Using the notation introduced in Section 3.4, each simple mass function  $m_k^Y(\cdot|A)$  can be denoted by  $\overline{\{y_k\}}^{-\log Pl_k^X(A)}$ .

3. Combine the  $n$  mass function  $\overline{\{y_k\}}^{-\log Pl_k^X(A)}$  using Dempster's rule. We get

$$m^Y(\cdot|A) = \bigoplus_{k=1}^n \overline{\{y_k\}}^{-\log Pl_k^X(A)}. \quad (6.26)$$

Equation (6.26) can be interpreted as follows. Assume that  $Pl_k^X(A)$  is low, which means that proposition  $X \in A$  is not very plausible given that  $Y = y_k$ . Then, the fact that  $X \in A$  supports the hypothesis that  $Y$  is not equal to  $y_k$ . In contrast, if  $Pl_k^X(A)$  is high, the mass function

$\overline{\{y_k\}}^{-\log Pl_k^X(A)}$  is almost vacuous: observing that  $X \in A$  does not support the hypothesis that  $Y$  equals  $y_k$  (because  $A$  may be plausible given other values  $y_\ell$  of  $Y$ , too). The same kind of reasoning is used in significance tests. Consider a significance test with p-value  $p = \mathbb{P}_{H_0}(T > t)$ , where  $H_0$  is the null hypothesis,  $T$  is the test statistic and  $t$  its realization. A small value of  $p$  is considered as evidence against  $H_0$ , but a high value of  $p$  is not considered to support  $H_0$ . In fact, a significance can only provide evidence against  $H_0$ , or fail to provide such evidence.

The degree of conflict between the  $n$  mass functions  $\overline{\{y_k\}}^{-\log Pl_k^X(A)}$  is

$$\kappa = \prod_{k=1}^n (1 - Pl_k^X(A)). \quad (6.27)$$

The conflict is thus high when  $A$  is implausible under all possible values of  $Y$ . The commonality function  $Q_k^Y(\cdot|A)$  corresponding to the mass function  $\overline{\{y_k\}}^{-\log Pl_k^X(A)}$  is

$$Q_k^Y(B|A) = \begin{cases} Pl_k^X(A) & \text{if } y \in B, \\ 1 & \text{otherwise,} \end{cases} \quad (6.28)$$

for all  $B \subseteq \Omega_Y$ . Consequently, the combined commonality function corresponding to (6.26) is

$$Q^Y(B|A) = K \prod_{y_k \in B} Pl_k^X(A), \quad (6.29)$$

where  $K = (1 - \kappa)^{-1}$ . Smets [89] shows that the corresponding plausibility function is:

$$Pl^Y(B|A) = K \left[ 1 - \prod_{y_k \in B} (1 - Pl_k^X(A)) \right]. \quad (6.30)$$

The GBT has some interesting properties. First, assume that the conditional mass functions  $m_k^X$  are Bayesian and we have some prior information on  $Y$  in the form of a Bayesian mass function  $m_0^Y$ . Then we can compute  $m^Y(\cdot|A)$  using (6.26) and combine it with the prior  $m_0^Y$ . The result is a Bayesian mass function defined as

$$m_1^Y(\{y_k\}|A) = \frac{m_k^X(A)m_0^Y(\{y_k\})}{\sum_{\ell=1}^n m_\ell^X(A)m_0^Y(\{y_\ell\})} \quad (6.31)$$

for all  $k$ . When provided with the same information, the GBT and the Bayes theorem thus lead to the same conclusions. However, the GBT does not require a prior (or, equivalently, the prior  $m_0^Y$  may be vacuous), and the conditional belief function on  $\Omega_X$  given  $y$  need not be Bayesian.

The second remarkable property is related to the notion of *cognitive independence* [77]. Assume that we have three variables  $X$ ,  $Y$  and  $Z$ . Variables  $X$  and  $Z$  are said to be *cognitively independent* conditionally on  $Y$  if, for all  $A \subseteq \Omega_X$ ,  $B \subseteq \Omega_Z$  and  $y_k \in \Omega_Y$ ,

$$pl^{XZ}(A \times B|y_k) = pl^X(A|y_k) \cdot pl^Z(B|y_k). \quad (6.32)$$

Assume that this property holds, and we observe  $X \in A$  and  $Z \in B$ . Then, we can equivalently apply the GBT to compute a conditional mass function  $m^Y(\cdot|X \in A, Z \in B)$  on  $\Omega_Y$  given the facts  $X \in A$  and  $Z \in B$  directly using (6.26), or we can compute  $m^Y(\cdot|X \in A)$  and  $m^Y(\cdot|Z \in B)$  separately and combine them using Dempster's rule:

$$m^Y(\cdot|X \in A, Z \in B) = m^Y(\cdot|X \in A) \oplus m^Y(\cdot|Z \in B). \quad (6.33)$$

## Exercises

1. Let  $\Theta = \{\theta_1, \theta_2\}$  and  $\Omega = \{a, b, c, d\}$  be two frames of discernment, and let  $\rho$  be the refinement defined by  $\rho(\{\theta_1\}) = \{a, b\}$  and  $\rho(\{\theta_2\}) = \{c, d\}$ .
  - (a) Compute the vacuous extension on  $\Omega$  of the mass function defined by  $m^\Theta(\emptyset) = 0.1$ ,  $m^\Theta(\{\theta_1\}) = 0.2$ ,  $m^\Theta(\{\theta_2\}) = 0.4$ ,  $m^\Theta(\Theta) = 0.3$ .
  - (b) Compute the restriction on  $\Theta$  of the mass function  $m^\Omega(\{a\}) = 0.1$ ,  $m^\Omega(\{a, b\}) = 0.2$ ,  $m^\Omega(\{a, b, c\}) = 0.3$ ,  $m^\Omega(\{c, d\}) = 0.2$ ,  $m^\Omega(\{b, d\}) = 0.2$ .
2. Let  $\Theta = \{\theta_1, \theta_2, \theta_3\}$  and  $\Omega = \{a, b, c\}$  be two frames of discernment. We consider the following mass function on  $\Omega \times \Theta$ :

$$m^{\Omega \times \Theta}(\{(a, \theta_1)\}) = 0.2 \quad m^{\Omega \times \Theta}(\Omega \times \{\theta_2\}) = 0.3$$

$$m^{\Omega \times \Theta}(\{b\} \times \Theta) = 0.4 \quad m^{\Omega \times \Theta}(\{(a, \theta_1), (b, \theta_2), (c, \theta_3)\}) = 0.1$$

- (a) Compute  $m^{\Omega \times \Theta \downarrow \Omega}$  and its vacuous extension on  $\Omega \times \Theta$ .
- (b) Compute  $m^{\Omega \times \Theta \downarrow \Theta}$  and its vacuous extension on  $\Omega \times \Theta$ .

## Chapter 7

# Decision-making

Quite often, if not always, the ultimate purpose of quantifying uncertainty is to make decisions. The general problem of decision-making under uncertainty has a long history and is of the utmost importance in many areas, such as economics and engineering. In this chapter, we will review different principles for making decisions, when uncertainty is quantified by belief functions. The formal framework will first be introduced in Section 7.1. The classical approaches to decision-making under complete ignorance, and in a context where uncertainty is probabilized, will then be described in Section 7.2. These preliminaries will set the ground for a review of the main approaches for decision-making with belief functions and their justifications, which will be presented in Section 7.3.

### 7.1 Formal framework

A decision problem can be seen as a situation in which a decision-maker (DM) has to choose a course of action (an *act*) in some set  $\mathcal{F}$ . An act may have different *consequences*, depending on the *state of nature*. Denoting by  $\Omega = \{\omega_1, \dots, \omega_n\}$  the set of states of nature and by  $\mathcal{C} = \{c_1, \dots, c_r\}$  the set of consequences (or outcomes), an act can thus be formalized as a mapping  $f$  from  $\Omega$  to  $\mathcal{C}$ . In most of this chapter, the three sets  $\Omega$ ,  $\mathcal{C}$  and  $\mathcal{F}$  will be assumed to be finite.

The desirability of the consequences can often be modeled by a *utility function*  $u : \mathcal{C} \rightarrow \mathbb{R}$ , which assigns a numerical value to each consequence. The higher this value, the more desirable is the consequence for the DM. In some problems, the consequences can be evaluated in terms of monetary value. The utilities can then be defined as the payoffs, or a function thereof.

Table 7.1: Payoff matrix (in €) for the investment example.

Act (Purchase)	Good Economic Conditions ( $\omega_1$ )	Poor Economic Conditions ( $\omega_2$ )
Apartment building ( $f_1$ )	50,000	30,000
Office building ( $f_2$ )	100,000	-40,000
Warehouse ( $f_3$ )	30,000	10,000

If the actions are indexed by  $i$  and the states of nature by  $j$ , we will denote by  $u_{ij}$  the quantity  $u[f_i(\omega_j)]$ . The  $n \times r$  matrix  $U = (u_{ij})$  will be called a payoff or utility matrix. These notions will now be illustrated using the following example taken from [70].

**Example 7.1** *Assume that an investor can decide between three acts: buying an apartment building ( $f_1$ ), an office building ( $f_2$ ), or a warehouse ( $f_3$ ). The consequences of these acts are payoffs, which depend on the economic conditions: good ( $\omega_1$ ) or poor ( $\omega_2$ ). Here, we assume the utilities to be equal to the payoffs. The utility matrix is shown in Table 7.1.*

If the true state of nature  $\omega$  is known, then the desirability of an act  $f$  can be deduced from that of its consequence  $f(\omega)$ . Typically, however, the state of nature is unknown. Based on partial information, it is usually assumed that the DM can express preferences among acts, which may be represented mathematically by a preference relation  $\succsim$  on  $\mathcal{F}$ . This relation is interpreted as follows: given two acts  $f$  and  $g$ ,  $f \succsim g$  means that  $f$  is found by the DM to be *at least as desirable* as  $g$ . We also define the strict preference relation as  $f \succ g$  iff  $f \succsim g$  and not( $g \succsim f$ ) (meaning that  $f$  is strictly more desirable than  $g$ ) and an indifference relation  $f \sim g$  iff  $f \succsim g$  and  $g \succsim f$  (meaning that  $f$  and  $g$  are equally desirable).

Quite often, the decision problem is to construct a preference relation among acts, from a utility matrix and some description of uncertainty, and to find the maximal elements of this relation. Depending on the nature of the available information, different decision problems arise. Several such problems will be considered in this chapter.

## 7.2 Elements of classical decision theory

In this section, we first consider the case where a utility function is given, but the DM is in a state of complete ignorance about the state of nature (Section 7.2.1). We will then consider the situation where probabilities on  $\Omega$  are given, which gives rise to the expected utility (EU) model (Section 7.2.2). Finally, in Section 7.2.3, we will discuss Savage's theorem, which states that, under uncertainty, a rational DM should rank acts according to EU, for some subjective probability measure and utility function.

### 7.2.1 Decision-making under complete ignorance

Let us start with the situation where the DM is totally ignorant of the state of nature. All the information given to the DM is thus the utility matrix  $U$ . An act  $f_i$  is said to be dominated by  $f_k$  if  $u_{ij} \leq u_{kj}$  for all  $j$ , and  $u_{ij} < u_{kj}$  for some  $j$ . It means that the consequences of act  $f_k$  are always at least as desirable as those of act  $f_i$ , whatever the state of nature. The non-domination principle [99] prohibits the choice of an act that is dominated by another one. For instance, in Table 7.1, we can see that act  $f_3$  is dominated by  $f_1$ : consequently, we can remove  $f_3$  from further consideration.

After all dominated acts have been removed, there remains the problem of ordering them by desirability, and of finding the set of most desirable acts. Several criteria of "rational choice" that have been proposed to derive a preference relation over acts. They are summarized in the following list (see, e.g., [59, 99]).

1. The *Laplace criterion* ranks acts according to the average utility of their consequences:  $f_i \succeq f_k$  iff

$$\frac{1}{n} \sum_j u_{ij} \geq \frac{1}{n} \sum_j u_{kj}. \quad (7.1)$$

2. The *maximax criterion* considers, for each act, its more favorable consequence. We then have  $f_i \succeq f_k$  iff

$$\max_j u_{ij} \geq \max_j u_{kj}. \quad (7.2)$$

3. Conversely, the *maximin criterion* takes into account the least favorable consequence of each act: act  $f_i$  is thus more desirable than  $f_k$  iff

$$\min_j u_{ij} \geq \min_j u_{kj}. \quad (7.3)$$

4. The *Hurwicz criterion* considers, for each act, a convex combination of the minimum and maximum utility:  $f_i \succeq f_k$  iff

$$\alpha \min_j u_{ij} + (1 - \alpha) \max_j u_{ij} \geq \alpha \min_j u_{kj} + (1 - \alpha) \max_j u_{kj}, \quad (7.4)$$

where  $\alpha$  is a parameter in  $[0, 1]$ , called the *pessimism index*.

5. Finally, the *minimax regret criterion* considers an act  $f_i$  to be at least as desirable than  $f_k$  if it has smaller maximal regret, where regret is defined as the utility difference with the best act, for a given state of nature; more precisely, let the regret  $r_{ij}$  for act  $f_i$  and state  $\omega_j$  be defined as follows,

$$r_{ij} = \max_{\ell} u_{\ell j} - u_{ij}. \quad (7.5)$$

The maximum regret for act  $f_i$  is thus  $R_i = \max_j r_{ij}$ , and act  $f_i$  is considered as more desirable than  $f_k$  when  $R_i \leq R_k$ .

**Example 7.2** Consider again Example 7.1. We have seen that act  $f_3$  is dominated and should be ruled out. It is easy to check that  $f_1 \succ f_2$  for the Laplace, maximin and minimax regret criteria, while  $f_2 \succ f_1$  for the maximax criterion. For the Hurwicz criterion, we let the reader verify that  $f_1 \succ f_2$  iff  $\alpha \geq 5/12$ .

Let us briefly comment on these criteria. The Laplace criterion is actually based on the expected utility, using a uniform probability distribution on the state of nature (see Section 7.2.2 below). It can thus be considered as resulting from the Principle of Indifference (see Section 1.3.5). The maximax and maximin criteria correspond, respectively, to extreme optimistic and pessimistic (or conservative) attitudes of the DM. The Hurwicz criterion allows to parameterize the DM's attitude toward ambiguity, using the pessimism index. These four criteria amount to extending the utility function to sets, i.e., they aggregate, for each act  $f_i$ , the utilities  $u_{ij}$  for all  $j$ , into a single number. The minimax regret criterion works differently, as it measures the desirability of an act by a quantity that depends on the consequences of all the other acts.

Given a set of acts  $\mathcal{F}$ , each of the five criteria listed above induces a complete preorder  $\succsim$  over the acts. Let  $\mathcal{F}^*$  denote the choice set, defined as the set of greatest elements of  $\succsim$ . Arrow and Hurwicz [6] have proposed four axioms a choice operator  $\mathcal{F} \rightarrow \mathcal{F}^*$  (i.e., a way of constructing the choice set) should verify. The following description of these axioms follows [36].



1. Axiom  $A_1$ : The non-empty intersection of a decision problem and the choice set of a larger decision problem is the choice set of the former. Formally, if  $\mathcal{F}_1 \subset \mathcal{F}_2$  and  $\mathcal{F}_2^* \cap \mathcal{F}_1 \neq \emptyset$ , then  $\mathcal{F}_1^* = \mathcal{F}_2^* \cap \mathcal{F}_1$ .
2. Axiom  $A_2$ : Relabeling actions and states does not change the optimal status of actions. Formally, if  $\phi_a$  is a one-to-one mapping from  $\mathcal{F}_1$  to  $\mathcal{F}_2$  and  $\phi_s$  is a one-to-one mapping from  $\Omega_1$  to  $\Omega_2$  such that, for all  $f \in \mathcal{F}_1$  and for all  $\omega \in \Omega_1$ ,  $f(\omega) = \phi_a(f)(\phi_s(\omega))$ , then  $f \in \mathcal{F}_1^*$  iff  $\phi_a(f) \in \mathcal{F}_2^*$ .

Given a set of actions  $\mathcal{F}$ , a state  $\omega \in \Omega$  is said to be duplicate if there exists another state  $\omega'$  in  $\Omega$  such that, for all  $f \in \mathcal{F}$ ,  $f(\omega) = f(\omega')$ . Deleting a duplicate state  $\omega$  means defining a new state space  $\Omega' = \Omega \setminus \{\omega\}$  and the new set of actions  $\mathcal{F}|_{\Omega'}$  containing the restrictions  $f|_{\Omega'}$  of all acts  $f$  in  $\mathcal{F}$ . We then have the following postulate.

3. Axiom  $A_3$ : Deletion of a duplicate state does not change the optimality status of actions. Formally,  $f \in \mathcal{F}^*$  iff  $f|_{\Omega'} \in (\mathcal{F}|_{\Omega'})^*$ .
4. Axiom  $A_4$  (dominance): If  $f \in \mathcal{F}^*$  and  $f'$  dominates  $f$ , then  $f' \in \mathcal{F}^*$ . If  $f \notin \mathcal{F}^*$  and  $f$  dominates  $f'$ , then  $f' \notin \mathcal{F}^*$ .

Axiom  $A_1$  is clearly violated by the minimax regret criterion. To see this, consider again, for instance, the data of Example 7.1. For  $\mathcal{F}_1 = \{f_1, f_2, f_3\}$ , we have seen that  $\mathcal{F}_1^* = \{f_1\}$  according to the minimax regret criterion. Now, consider a new act  $f_4$  such that  $u_{41} = 130,000$  and  $u_{42} = -45,000$ , and  $\mathcal{F}_2 = \{f_1, f_2, f_3, f_4\}$ . As a consequence of the introduction of this new act, the maximal regrets of  $f_1, f_2, f_3$  and  $f_4$  now become, respectively, 80,000, 70,000, 100,000 and 75,000. Hence,  $\mathcal{F}_2^* = \{f_2\}$  and  $\mathcal{F}_1^* \neq \mathcal{F}_2^* \cap \mathcal{F}_1$ .

It is also easy to see that Axiom  $A_3$  is violated by the Laplace criterion. To illustrate this point, assume that, in Example 7.1, we split the state of nature  $\omega_1$  in two states: “Good economic conditions and there is life on Mars” ( $\omega'_1$ ) and “Good economic conditions and there is no life on Mars” ( $\omega''_1$ ). It is clear that the payoffs of the real estate investments are identical under  $\omega'_1$  and  $\omega''_1$ . Consequently, the new payoff matrix will be obtained by duplicating the first column in Table 7.1. With this payoff matrix and  $\mathcal{F} = \{f_1, f_2, f_3\}$ , the choice set is  $\mathcal{F}^* = \{f_2\}$ . Should be learn that there is no life of Mars, we would delete  $\omega'_1$  and get that same payoff matrix as in Table 7.1. The corresponding choice set is now the restriction of  $f_2$  to  $\{\omega''_1, \omega_2\}$ , i.e., after learning that there is no life on Mars, we now change our investment decision!

It thus seems that the Laplace and minimax regret criteria must be ruled out as criteria for rational decision-making under ignorance. A stronger result, due to Arrow and Hurwicz [6] is that, under some regularity assumptions, Axioms  $A_1 - A_4$  imply that the choice set depends only on the worst and the best consequences of each act. This finding led to the proposal of the Hurwicz criterion (7.4).

For completeness, we should also mention in this section another criterion for decision-making under ignorance, recently introduced by Yager [112], which extends the Laplace, maximax, maximin and Hurwicz criteria. Each of these criteria corresponds to a different way of aggregating the utilities resulting each act, using, respectively, the average, the maximum, the minimum, and a convex sum of the minimum and the maximum. These four operators happen to belong to a family of operators called Ordered Weighted Average (OWA) operators [112]. An OWA operator of dimension  $n$  is a function  $F : \mathbb{R}^n \rightarrow \mathbb{R}$  of the form

$$F(x_1, \dots, x_n) = \sum_{i=1}^n w_i x_{(i)}, \quad (7.6)$$

where  $x_{(i)}$  is the  $i$ -th largest element in the collection  $x_1, \dots, x_n$ , and  $w_1, \dots, w_n$  are positive weights verifying

$$\sum_{i=1}^n w_i = 1. \quad (7.7)$$

It is clear that the four above-mentioned operators are obtained for different choices of the weights:

- Average:  $(1/n, 1/n, \dots, 1/n)$ ;
- Maximum:  $(1, 0, \dots, 0)$ ;
- Minimum:  $(0, \dots, 0, 1)$ ;
- Hurwicz:  $(1 - \alpha, 0, \dots, 0, \alpha)$ .

In a decision-making context, each weight  $w_i$  may be interpreted as a probability that the  $i$ -th best outcome will happen. Yager [112] defines the degree of optimism of an OWA operator with weight vector  $\mathbf{w}$  as

$$OPT(\mathbf{w}) = \sum_{i=1}^n \frac{n-i}{n-1} w_i. \quad (7.8)$$

Given a degree of optimism  $\beta$ , we can then choose the OWA operator that maximizes the entropy

$$ENT(\mathbf{w}) = - \sum_{i=1}^n w_i \log w_i, \quad (7.9)$$

under the constraint  $OPT(\mathbf{w}) = \beta$ .

### 7.2.2 Decision-making with probabilities

Let us now consider the situation where uncertainty about the state of nature is quantified by probabilities  $p_1, \dots, p_n$  on  $\Omega$ . Typically, these probabilities are assumed to be objective: we say that we have a problem on *decision under risk*. However, the following developments also apply to the case where the probabilities are subjective. In any case, the probability distribution  $p_1, \dots, p_n$  is assumed to be known, together with the utility matrix  $U$ . We can then compute, for each act  $f_i$ , its expected utility as

$$EU(f_i) = \sum_j u_{ij} p_j. \quad (7.10)$$

According to the Maximum Expected Utility (MEU) principle, an act  $f_i$  is more desirable than an act  $f_k$  if it yields more desirable consequences *on average* over all possible states of nature, i.e., if it has a higher expected utility:  $f_i \succeq f_k$  iff  $EU(f_i) \geq EU(f_k)$ .

**Example 7.3** *Continuing Examples 7.1 and 7.2, assume that there is 60% chance that the economic situation will be poor. The expected utilities of acts  $f_1$  and  $f_2$  are*

$$EU(f_1) = 50,000 \times 0.4 + 30,000 \times 0.6 = 38,000 \quad (7.11a)$$

$$EU(f_2) = 100,000 \times 0.4 - 40,000 \times 0.6 = 16,000. \quad (7.11b)$$

*Act  $f_1$  is thus clearly more desirable according to the maximum expected utility criterion.*

The MEU principle was first axiomatized by von Neumann and Morgenstern [103]. We give hereafter a summary of their argument. Given a probability distribution on  $\Omega$ , an act  $f : \Omega \rightarrow \mathcal{C}$  induces a probability measure  $P$  on the set  $\mathcal{C}$  of consequences (assumed to be finite), called a *lottery*. We denote by  $\mathcal{L}$  the set of lotteries on  $\mathcal{C}$ . If we agree that two acts providing

the same lottery are equivalent, then the problem of comparing the desirability of acts becomes that of comparing the desirability of lotteries. Let  $\succeq$  be a preference relation among lotteries. Von Neumann and Morgenstern argued that, to be rational, a preference relation should verify the following three axioms.

1. *Complete preorder*: the preference relation is a complete and non trivial preorder (i.e., it is a reflexive, transitive and complete relation) on  $\mathcal{L}$ .
2. *Continuity*: for any lotteries  $P$ ,  $Q$  and  $R$  such that  $P \succ Q \succ R$ , there exists a probabilities  $\alpha$  and  $\beta$  in  $[0, 1]$  such that

$$\alpha P + (1 - \alpha)R \succ Q \succ \beta P + (1 - \beta)R \quad (7.12)$$

where  $\alpha P + (1 - \alpha)R$  is a compound lottery, which refers to the situation where you receive  $P$  with probability  $\alpha$  and  $Q$  with probability  $1 - \alpha$ . This axiom implies, in particular, that there is no lottery  $R$  that is so undesirable that it cannot become desirable if mixed with some very desirable lottery  $P$ .

3. *Independence*: for any lotteries  $P$ ,  $Q$  and  $R$  and for any  $\alpha \in (0, 1]$ ,

$$P \succeq Q \Leftrightarrow \alpha P + (1 - \alpha)R \succeq \alpha Q + (1 - \alpha)R. \quad (7.13)$$

We then have the following theorem.

**Theorem 7.1 (Von Neumann and Morgenstern)** *The two following propositions are equivalent:*

1. *The preference relation  $\succeq$  verifies the axioms of complete preorder, continuity, and independence;*
2. *There exists a utility function  $u : \mathcal{C} \rightarrow \mathbb{R}$  such that, for any two lotteries  $P = (p_1, \dots, p_r)$  and  $Q = (q_1, \dots, q_r)$ ,*

$$P \succeq Q \Leftrightarrow \sum_{i=1}^r p_i u(c_i) \geq \sum_{i=1}^r q_i u(c_i). \quad (7.14)$$

*Function  $u$  is unique up to a strictly increasing affine transformation.*

Von Neumann and Morgenstern's theorem has had a tremendous impact, as it provides a compelling justification of both the notion of utility, and the MEU principle. For problems of decision under risk, the normative value of the MEU principle is widely accepted. From a descriptive point of view, violations of the MEU principle by most DMs in some particular situations have been demonstrated experimentally. The first such experiment was reported by Allais [3], who proposed subjects to choose between the following two lotteries:

- $L_1$ : win 1M for sure;
- $L_2$ : win 1M with probability 0.89, 5M with probability 0.10 and 0 with probability 0.01.

The same subjects were then asked to choose between the following other two lotteries;

- $L_3$ : win 1M with probability 0.11 and 0 with probability 0.89;
- $L_4$ : win 5M with probability 0.10 and 0 with probability 0.90.

Most subjects strictly prefer  $L_1$  to  $L_2$ , and  $L_4$  to  $L_3$ . These choices violate the axiom of independence. To see this, let us denote by  $P$  the lottery that gives 1M with probability 1 and by  $Q$  the lottery that gives 5M with probability 1/11 and 0 with probability 10/11. We can check that

$$L_1 = 0.11P + 0.89\delta_1 \quad (7.15a)$$

$$L_2 = 0.11Q + 0.89\delta_1 \quad (7.15b)$$

$$L_3 = 0.11P + 0.89\delta_0 \quad (7.15c)$$

$$L_4 = 0.11Q + 0.89\delta_0, \quad (7.15d)$$

where  $\delta_1$  is the lottery that yields 1M for sure, and  $\delta_0$  is the lottery that yields 0 for sure. The observed choices thus contract the axiom of independence.

### 7.2.3 Savage's theorem

In Section 7.2.1, we reviewed some criteria for decision-making under complete ignorance, i.e., when uncertainty cannot be probabilized. Some researchers, however, have attempted to argue that a rational DM should always maximize expected utility, for some subjective probability measure and utility function. The first such attempt was made by Savage [74], who provided an axiomatic justification of the MEU principle for decision under

uncertainty. Savage proposed seven axioms that a preference relation among acts should verify. We will only discuss here the first four of these axioms, the other three being essentially technical.

The first axiom (ordering) states that  $\succsim$  should be a complete, reflexive and transitive relation. In particular, completeness implies that the DM is always able to choose between any two acts, even in the absence of any evidence about  $\Omega$ .

To introduce the second axiom, we need the following definition. Given two acts  $f, h \in \mathcal{F}$  and a subset  $E$  of  $\Omega$ , let  $fEh$  denote the act defined by

$$(fEh)(\omega) = \begin{cases} f(\omega) & \text{if } \omega \in E \\ h(\omega) & \text{if } \omega \notin E. \end{cases} \quad (7.16)$$

The act  $fEh$  thus has the same consequence as  $f$  if  $E$  is true, and the same consequences as  $h$  otherwise. The second axiom, called the the Sure Thing Principle (STP), states that for any event  $E \subseteq \Omega$  and any acts  $f, g, h, h'$ ,

$$fEh \succsim gEh \Rightarrow fEh' \succsim gEh'. \quad (7.17)$$

In other words, the preference between two acts with a common extension outside some event  $E$  does not depend on this common extension. This axiom seems reasonable. However, we will see that it is not verified experimentally.

Without any risk of ambiguity, let us use the same notation for any outcome  $x \in \mathcal{C}$  and the constant act such that  $f(\omega) = x$  for all  $\omega \in \Omega$ . According to the third axiom (monotonicity), for any event  $E$  and for any two constant acts  $x$  and  $y$ , if  $x$  is preferred to  $y$ , it is so whatever the state in which these outcomes are obtained:

$$x \succ y \Leftrightarrow (\forall f \in \mathcal{F}, xEf \succ yEf). \quad (7.18)$$

Lastly, the fourth axiom (Weak comparative ordering) can be stated as follows: for all events  $A$  and  $B$  and outcomes  $x' \succ x$  and  $y' \succ y$ ,

$$x'Ax \succ xBx' \Rightarrow y'Ay \succ yBy'. \quad (7.19)$$

Since  $x'$  is strictly preferred to  $x$ , the first preference rank means that the DM considers  $A$  to be more likely than  $B$ . This judgement should not depend on the particular outcomes  $x' \succ x$ , which is the meaning of this axiom.

Savage showed that any preference relation  $\succsim$  satisfies the above four axioms (plus three more technical ones) iff there is a finitely additive probability measure  $P$  and a utility function  $u : \mathcal{C} \rightarrow \mathbb{R}$  such that  $f$  is at most

preferable to  $g$  iff the expected utility of  $f$  is not smaller than that of  $g$ ,

$$\forall f, g \in \mathcal{F}, \quad f \succcurlyeq g \Leftrightarrow \int_{\Omega} u(f(\omega))dP(\omega) \geq \int_{\Omega} u(g(\omega))dP(\omega). \quad (7.20)$$

Moreover,  $P$  is unique and  $u$  is unique up to a positive affine transformation.

Savage's axioms imply that the state of nature is infinite. However, a similar result has been obtained for a finite number of states [39]. Savage's important result was first considered by many researchers to justify both subjective probabilities as the only rational representation of uncertainty, and the maximum utility decision criterion. However, the relevance of Savage's axioms can be questioned (see, e.g., [81]). In particular, it can be shown experimentally that the way people make decisions in the presence of ambiguity is not always consistent with the STP, as shown in the next section.

#### 7.2.4 Ellsberg's paradox

The so-called "Ellsberg's paradox" [32] is actually not a paradox, but an experimental demonstration that most people are not expected utility maximizers. Suppose you have an urn containing 30 red balls and 60 balls, either black or yellow. You are given a choice between two gambles:

- $f_1$ : You receive 100 euros if you draw a red ball;
- $f_2$ : You receive 100 euros if you draw a black ball.

Also, you are given a choice between these two gambles (about a different draw from the same urn):

- $f_3$ : You receive 100 euros if you draw a red or yellow ball;
- $f_4$ : You receive 100 euros if you draw a black or yellow ball.

Most people strictly prefer  $f_1$  to  $f_2$ , but they strictly prefer  $f_4$  to  $f_3$ . Is this behavior compatible with expected utility maximization? The outcomes for each of the four acts  $f_1$ ,  $f_2$ ,  $f_3$  and  $f_4$  are shown in Table 7.2. The expected utilities are

$$\begin{aligned} u(f_1) &= u(100)P(\text{red}) + u(0)P(\text{black}) + u(0)P(\text{yellow}), \\ u(f_2) &= u(0)P(\text{red}) + u(100)P(\text{black}) + u(0)P(\text{yellow}), \\ u(f_3) &= u(100)P(\text{red}) + u(0)P(\text{black}) + u(100)P(\text{yellow}), \\ u(f_4) &= u(0)P(\text{red}) + u(100)P(\text{black}) + u(100)P(\text{yellow}). \end{aligned}$$

Table 7.2: Outcomes in Ellsberg's paradox.

Act	red	black	yellow
$f_1$	100	0	0
$f_2$	0	100	0
$f_3$	100	0	100
$f_4$	0	100	100

From  $f_1 \succ f_2$ , we can deduce that  $P(\text{red}) > P(\text{black})$ , while  $f_4 \succ f_3$  entails that  $P(\text{black}) > P(\text{red})$ . The conclusion is people's decision cannot be explained by the MEU principle. As closer inspection reveals, this phenomenon is due to a violation of the STP. Indeed, the acts  $f_1$  and  $f_2$  coincide, respectively, with  $f_3$  and  $f_4$  when the outcome is "red" or "black". We can write:

$$f_1 = f_1\{\text{red, black}\}0 \qquad f_2 = f_2\{\text{red, black}\}0 \qquad (7.21a)$$

$$f_3 = f_1\{\text{red, black}\}100 \qquad f_4 = f_2\{\text{red, black}\}100. \qquad (7.21b)$$

It is clear that the most frequent observed preferences, i.e.,  $f_1 \succ f_2$  and  $f_4 \succ f_3$ , violate the STP.

### 7.3 Decision-making with belief functions

After the rather long review of classical decision theory presented in the previous section, we are now ready to tackle our main topic in this chapter, namely: decision-making in the belief function framework. Since, as repeatedly emphasized all along this book, belief functions combine sets with probabilities, it will come as no surprise that decision criteria with belief functions somehow mix the decision criteria under ignorance reviewed in Section 7.2.1, with the MEU criterion discussed in Section 7.2.2.

Belief functions become of component of a decision problem in the following two situations:

1. The decision maker's subjective beliefs concerning the possible states of nature may be described by a belief function  $Bel^\Omega$  on  $\Omega$ . Any act  $f$  then carries  $Bel^\Omega$  to the set  $\mathcal{C}$  of consequences, yielding a belief function  $Bel^\mathcal{C}$  defined as

$$Bel^\mathcal{C}(B) = Bel^\Omega(f^{-1}(B)), \qquad (7.22)$$



for any  $B \subseteq \mathcal{C}$ . The corresponding mass function is

$$m^{\mathcal{C}}(B) = \sum_{\{A \subseteq \Omega | f(A)=B\}} m^{\Omega}(A). \quad (7.23)$$

for any  $B \subseteq \mathcal{C}$ .

2. The consequences of each act under each state of nature may not be precisely described. As discussed in [35], this situation may arise when the decision problem is underspecified: for instance, the choice set (set of acts) or the state space might be too coarsely defined. In that case, an act may formally be represented by a multi-valued mapping  $f : \Omega \rightarrow 2^{\mathcal{C}}$ , assigning a set of possible consequences  $f(\omega) \subseteq \mathcal{C}$  to each state of nature  $\omega$ . Given a probability measure  $P$  on  $\Omega$ ,  $f$  then induces the following mass function  $m^{\mathcal{C}}$  on  $\mathcal{C}$ ,

$$m^{\mathcal{C}}(B) = \sum_{\{\omega \in \Omega | f(\omega)=B\}} p(\omega). \quad (7.24)$$

for all  $B \subseteq \Omega$ .

It is clear that these two situations can occur simultaneously, i.e., we may have a belief function  $Bel$  on  $\omega$ , and ill-known consequences. In that case, Equations (7.23) and (7.24) become

$$m^{\mathcal{C}}(B) = \sum_{\{A \subseteq \Omega | \bigcup_{\omega \in A} f(\omega)=B\}} m^{\Omega}(A). \quad (7.25)$$

for all  $B \subseteq \Omega$ .

**Example 7.4** Let  $\Omega = \{\omega_1, \omega_2, \omega_3\}$  and  $m^{\Omega}$  the following mass function on  $\Omega$ ,

$$\begin{aligned} m^{\Omega}(\{\omega_1, \omega_2\}) &= 0.3, & m^{\Omega}(\{\omega_2, \omega_3\}) &= 0.2, \\ m^{\Omega}(\{\omega_3\}) &= 0.4, & m^{\Omega}(\Omega) &= 0.1. \end{aligned} \quad (7.26)$$

Let  $\mathcal{C} = \{c_1, c_2, c_3\}$  and  $f$  the act

$$f(\omega_1) = \{c_1\}, \quad f(\omega_2) = \{c_1, c_2\}, \quad f(\omega_3) = \{c_2, c_3\}. \quad (7.27)$$

To compute the induced mass function on  $\mathcal{C}$ , we transfer the masses as follows,

$$m^{\Omega}(\{\omega_1, \omega_2\}) = 0.3 \rightarrow f(\omega_1) \cup f(\omega_2) = \{c_1, c_2\} \quad (7.28a)$$

$$m^{\Omega}(\{\omega_2, \omega_3\}) = 0.2 \rightarrow f(\omega_2) \cup f(\omega_3) = \{c_1, c_2, c_3\} \quad (7.28b)$$

$$m^{\Omega}(\{\omega_3\}) = 0.4 \rightarrow f(\omega_3) = \{c_2, c_3\} \quad (7.28c)$$

$$m^{\Omega}(\Omega) = 0.1 \rightarrow f(\omega_1) \cup f(\omega_2) \cup f(\omega_3) = \{c_1, c_2, c_3\}. \quad (7.28d)$$

Finally, we obtain the following mass function on  $\mathcal{C}$ ,

$$m^{\mathcal{C}}(\{c_1, c_2\}) = 0.3, \quad m^{\mathcal{C}}(\{c_2, c_3\}) = 0.4, \quad m^{\mathcal{C}}(\mathcal{C}) = 0.3. \quad (7.29)$$

□

In any of the situations considered above, we can assign to each act  $f$  a “credibilistic lottery”, defined as a mass function on  $\mathcal{C}$ . Given a utility function  $u$  on  $\mathcal{C}$ , we then need to extend the EU model. Several such extensions are briefly reviewed in the sequel.

### 7.3.1 Upper and lower expected utility

Let  $m$  be a mass function<sup>1</sup> on  $\mathcal{C}$ , and  $u$  a utility function  $\mathcal{C} \rightarrow \mathbb{R}$ . The lower and upper expectations of  $u$  are defined, respectively, as the averages of the minima and the maxima of  $u$  within each focal set of  $m$  [18, 79],

$$\underline{\mathbb{E}}_m(u) = \sum_{A \subseteq \mathcal{C}} m(A) \min_{c \in A} u(c), \quad (7.30a)$$

$$\overline{\mathbb{E}}_m(u) = \sum_{A \subseteq \mathcal{C}} m(A) \max_{c \in A} u(c). \quad (7.30b)$$

It is clear that  $\underline{\mathbb{E}}_m(u) \leq \overline{\mathbb{E}}_m(u)$ , with the inequality becoming an equality when  $m$  is Bayesian, in which case the lower and upper expectations collapse to the usual expectation. If  $m$  is logical with focal set  $A$ , then  $\underline{\mathbb{E}}_m(u)$  and  $\overline{\mathbb{E}}_m(u)$  are, respectively, the minimum and the maximum of  $u$  in  $A$ .

It can be shown [38] that the lower and upper expectations are Choquet integrals [13] with respect the belief and plausibility functions, respectively (see the appendix at the end of this chapter). Furthermore, they can be interpreted as lower and upper bounds of expectations with respect to probability measures compatible with  $Bel$  [79, 106, 38]. In other words, the mean of minima in (7.30a) is also the minimum of means with respect to all compatible probability measures; similarly, the mean of maxima (7.30b) is the maximum of means [38]. We recall that a probability measure  $P$  is compatible with  $Bel$  iff  $P \geq Bel$ . The set of such probability measures is called the credal set of  $Bel$  and is denoted by  $\mathcal{P}(Bel)$  (see Section 2.3.3). Now, for any nonempty subset  $A$  of  $\mathcal{C}$ , let  $c_A$  denote an element of  $A$  with minimum value of  $u$ , i.e., such that  $u(c_A) = \min_{c \in \mathcal{C}} u(c)$ . Let  $p_*$  be the probability

<sup>1</sup>In this and the following sections, all mass functions will be defined on  $\mathcal{C}$ , unless otherwise specified. For this reason, the notation  $m^{\mathcal{C}}$  will be simplified to  $m$ , without any risk of confusion.

distribution computed from  $m$  using the allocation  $\alpha_*$  that transfers each mass  $m(A)$  to  $c_A$ :

$$\alpha_*(c, A) = \begin{cases} m(A) & \text{if } c = c_A \\ 0 & \text{otherwise.} \end{cases} \quad (7.31)$$

We have

$$\mathbb{E}_{p_*}(u) = \sum_{c \in \mathcal{C}} p_*(c)u(c) = \sum_{A \subseteq \mathcal{C}} m(A)u(c_A) = \underline{\mathbb{E}}_m(u). \quad (7.32)$$

It is clear that  $\mathbb{E}_{p_*}(u) \leq \mathbb{E}_{p_\alpha}(u)$ , for all  $p_\alpha \in \mathcal{P}(Bel)$ . Consequently,

$$\underline{\mathbb{E}}_m(u) = \min_{P \in \mathcal{P}(Bel)} \mathbb{E}_P(u). \quad (7.33)$$

Using a similar line of reasoning, it can easily be shown that

$$\overline{\mathbb{E}}_m(u) = \max_{P \in \mathcal{P}(Bel)} \mathbb{E}_P(u). \quad (7.34)$$

The interval  $[\underline{\mathbb{E}}_m(u), \overline{\mathbb{E}}_m(u)]$  can thus be seen as the range of  $\mathbb{E}_P(u)$  for all probability measures  $P$  compatible with  $Bel$ .

Having defined the notions of lower and upper expectations, we can define two preference relations among credibilistic lotteries as,

$$m_1 \succeq m_2 \text{ iff } \underline{\mathbb{E}}_{m_1}(u) \geq \underline{\mathbb{E}}_{m_2}(u), \quad (7.35)$$

and

$$m_1 \overline{\succeq} m_2 \text{ iff } \overline{\mathbb{E}}_{m_1}(u) \geq \overline{\mathbb{E}}_{m_2}(u). \quad (7.36)$$

Relation  $\succeq$  corresponds to a pessimistic (or conservative) attitude of the DM, since it takes in account the least favorable consequence within each focal set. When  $m$  is logical, then  $\succeq$  corresponds to the maximin criterion. Symmetrically,  $\overline{\succeq}$  corresponds to an optimistic attitude and extends the maximax criterion. Both criteria boil down to the EU criterion when  $m$  is Bayesian.

**Example 7.5** *Let us come back to Ellsberg's paradox. Knowing that the urn contains 30 red balls and 60 balls, either black or yellow, a natural mass function about the color of a ball to be drawn from the urn is  $m^\Omega(\{R\}) = 1/3$ ,  $m^\Omega(\{B, Y\}) = 2/3$ , where  $\Omega = \{R, B, Y\}$  is the frame of discernment. The mass functions on  $\mathcal{C} = \{0, 100\}$  induced by the four acts  $f_1$  to  $f_4$  are*

$$m_1(\{100\}) = 1/3, \quad m_1(\{0\}) = 2/3, \quad (7.37a)$$

$$m_2(\{0\}) = 1/3, \quad m_2(\{0, 100\}) = 2/3, \quad (7.37b)$$

$$m_3(\{100\}) = 1/3, \quad m_3(\{0, 100\}) = 2/3, \quad (7.37c)$$

$$m_4(\{0\}) = 1/3, \quad m_4(\{100\}) = 2/3. \quad (7.37d)$$

Table 7.3: Lower and upper expected utilities in Ellsberg's urn experiment.

	$R$	$B$	$Y$	$\underline{\mathbb{E}}_m(u)$	$\overline{\mathbb{E}}_m(u)$
$f_1$	100	0	0	$u(100)/3$	$u(100)/3$
$f_2$	0	100	0	$u(0)$	$2u(100)/3$
$f_3$	100	0	100	$u(100)/3$	$u(100)$
$f_4$	0	100	100	$2u(100)/3$	$2u(100)/3$

The lower and upper expectations of  $u$  with respect to each of these four credibilistic lotteries are shown in Table 7.3. We can see that the observed behavior ( $f_1 \succ f_2$  and  $f_4 \succ f_3$ ) is explained by the pessimistic strategy.  $\square$

If one drops the requirement that the preference relation among acts be complete, then one can adopt the following partial preference relation, called the *interval dominance* relation,

$$m_1 \succ_{ID} m_2 \text{ iff } \underline{\mathbb{E}}_{m_1}(u) \geq \overline{\mathbb{E}}_{m_2}(u). \quad (7.38)$$

Given a collection of credibilistic lotteries, we can then consider the set of maximal (i.e., non dominated) elements of  $\succ_{ID}$ . From an imprecise probability perspective [101],  $m_1 \succ_{ID} m_2$  iff the expectation of  $u$  with respect to a probability measure compatible with  $m_1$  is always higher than the expectation of  $u$  with respect to a probability measure compatible with  $m_2$ ,

$$m_1 \succ_{ID} m_2 \Leftrightarrow \forall P_1 \in \mathcal{P}(m_1), \forall P_2 \in \mathcal{P}(m_2), \mathbb{E}_{P_1}(u) \geq \mathbb{E}_{P_2}(u). \quad (7.39)$$

The justification for this preference relation, however, is not so clear from the point of view of belief function theory (i.e., if one does not interpret a belief function as a lower probability).

### 7.3.2 Other approaches

Just as the lower and upper expected utility models generalize, respectively, the maximin and maximax criteria, the Hurwicz criterion (7.4) can be readily generalized by defining the expectation of  $u$ , for a pessimism index  $\alpha \in [0, 1]$ , as

$$\mathbb{E}_{m,\alpha}(u) = \sum_{A \subseteq \mathcal{C}} m(A) \left( \alpha \min_{c \in A} u(c) + (1 - \alpha) \max_{c \in A} u(c) \right) \quad (7.40a)$$

$$= \alpha \underline{\mathbb{E}}_m(u) + (1 - \alpha) \overline{\mathbb{E}}_m(u). \quad (7.40b)$$

This criterion was introduced by Jaffray [43, 44], who also justified it axiomatically (see Section 7.3.3 below). It was discussed by Strat [98], who proposed to interpret  $\alpha$  as the DM's subjective probability that the ambiguity will be resolved unfavorably (see also [97] for a discussion of this criterion). Hereafter, we will use the term “Hurwicz criterion” to refer to decision based on (7.40), as this principle is a direct extension to the Hurwicz criterion in the case of complete ignorance.

A completely different approach to decision-making with belief function was advocated by Smets, as part of the Transferable Belief Model [88, 95, 92, 93]. Smets defended a two-level mental model, composed of a credal level, where an agent's belief are represented by belief functions, and the pignistic level, where decisions are made by maximizing EU with respect to a probability measure derived from a belief function. The rationale for introducing probabilities at the decision level is the avoidance of Dutch books (see Section 1.3.2). Furthermore, Smets argued that the belief-probability transformation  $T$  should be linear, i.e., it should verify

$$T(\alpha m_1 + (1 - \alpha)m_2) = \alpha T(m_1) + (1 - \alpha)T(m_2), \quad (7.41)$$

for any mass functions  $m_1$  and  $m_2$  and for any  $\alpha \in [0, 1]$ . He then showed that the only transformation  $T$  verifying (7.41) is the *pignistic transformation*, with  $p_m = T(m)$  given by

$$p_m(c) = \sum_{\{A \subseteq \mathcal{C} | c \in A\}} \frac{m(A)}{|A|}, \quad (7.42)$$

for any  $c \in \mathcal{C}$ . The pignistic probability  $p_m$  turns out to be mathematically identical to the Shapley value in cooperative game theory [84]. The expected utility w.r.t. the pignistic probability is

$$\mathbb{E}_p(u) = \sum_{c \in \mathcal{C}} p_m(c) u(c) \quad (7.43a)$$

$$= \sum_{c \in \mathcal{C}} u(c) \sum_{\{A \subseteq \mathcal{C} | c \in A\}} \frac{m(A)}{|A|} \quad (7.43b)$$

$$= \sum_{\{A \subseteq \mathcal{C}\}} m(A) \left( \frac{1}{|A|} \sum_{c \in A} u(c) \right). \quad (7.43c)$$

The maximum pignistic expected utility criterion thus extends the Laplace criterion introduced in Section 7.2.1, when uncertainty is quantified by a belief function.

A more general family of expected utility criteria can be defined by aggregating the utilities  $u(c)$  within each focal set  $A$  using OWA operators (7.6) [113]. It is clear that the previous definitions in this section are recovered as special cases. To determine the weights of the OWA operators, Yager [113] proposes to fix the degree of optimism  $\beta$  defined by (7.8), and to use the maximum-entropy operators, for each cardinality  $|A|$ . Formally,

$$\mathbb{E}_{m,\beta}^{\text{owa}} = \sum_{A \subseteq \mathcal{C}} m(A) F_{|A|,\beta}(\{u(c) | c \in A\}), \quad (7.44)$$

where  $F_{|A|,\beta}$  is the maximum-entropy OWA operator with degree of optimism  $\beta$  and arity  $|A|$ . We can remark that parameter  $\beta$  plays the same role here, and has roughly the same interpretation, as one minus the pessimism index  $\alpha$  in the Hurwicz criterion. However, each quantity  $F_{|A|,\beta}(\{u(c) | c \in A\})$  depends on all the values  $u(c)$  for all  $c \in A$ , and not only on the minimum and the maximum. This method is further discussed in [108].

Finally, Yager [114] also extended the minimax regret criterion to belief functions. As in Section 7.2.1, we need to consider  $n$  acts  $f_1, \dots, f_n$ , and we write  $u_{ij} = u[f_i(\omega_j)]$ . The regret if act  $f_i$  is selected, and state  $\omega_j$  occurs, is  $r_{ij} = \max_k u_{kj} - u_{ij}$ . For a non-empty subset  $A$  of  $\Omega$ , the maximum regret of act  $f_i$  is

$$R_i(A) = \max_{\omega_j \in A} r_{ij}. \quad (7.45)$$

Given a mass function  $m^\Omega$  on  $\Omega$ , the expected maximal regret for act  $f_i$  is

$$\bar{R}_i = \sum_{\emptyset \neq A \subseteq \Omega} m^\Omega(A) R_i(A). \quad (7.46)$$

Using the generalized minimax regret criterion, act  $f_i$  is preferred over act  $f_k$  if  $\bar{R}_i \leq \bar{R}_k$ . The minimax regret criterion of decision-making under ignorance is recovered when  $m^\Omega$  is logical. If  $m^\Omega$  is Bayesian, we have

$$\bar{R}_i = \sum_j m^\Omega(\{\omega_j\}) r_{ij} \quad (7.47a)$$

$$= \sum_j m^\Omega(\{\omega_j\}) (\max_k u_{kj} - u_{ij}) \quad (7.47b)$$

$$= \sum_j m^\Omega(\{\omega_j\}) \max_k u_{kj} - \sum_j m^\Omega(\{\omega_j\}) u_{ij}. \quad (7.47c)$$

The first term on the righthand side of the last equation does not depend on  $i$ , and the second one is the expected utility. Hence, the generalized minimax regret criterion is identical with the EU model when  $m^\Omega$  is Bayesian.

### 7.3.3 Axiomatic justifications

Except for the generalized minimax regret criterion, the decision criteria reviewed in Sections 7.3.1 and 7.3.2 above are all of the form

$$m_1 \succcurlyeq m_2 \text{ iff } U(m_1) \geq U(m_2), \quad (7.48)$$

where  $U$  is a function from the set of mass functions on  $\mathcal{C}$  to  $\mathbb{R}$ , such that

$$U(m) = \sum_{\emptyset \neq A \subseteq \mathcal{C}} m(A)U(m_A), \quad (7.49)$$

where  $m_A$  is the logical mass function with focal set  $A$ . To simplify the notation, we can write  $U(A)$  in place of  $U(m_A)$ , and  $u(c)$  for  $U(\{c\})$ . With these notations, we have

- $U(A) = \min_{c \in A} u(c)$  for the lower expectation criterion;
- $U(A) = \max_{c \in A} u(c)$  for the upper expectation criterion;
- $U(A) = \alpha \min_{c \in A} u(c) + (1 - \alpha) \max_{c \in A} u(c)$  for the Hurwicz criterion;
- $U(A) = (1/|A|) \sum_{c \in A} u(c)$  for the pignistic criterion;
- $U(A) = F_{|A|, \beta}(\{u(c) | c \in A\})$  for the OWA criterion.

Jaffray [44] showed that a preference relation among credibilistic lotteries is representable by a linear utility function verifying (7.49) if and only if it verifies the Von Neumann and Morgenstern axioms (see Section 7.2.2) extended to credibilistic lotteries, i.e.,

1. Transitivity and Completeness:  $\succcurlyeq$  is a transitive and complete relation (i.e., is a weak order);
2. Continuity: for all  $m_1, m_2$  and  $m_3$  such that  $m_1 \succ m_2 \succ m_3$ , there exists  $\alpha, \beta$  in  $(0, 1)$  such that

$$\alpha m_1 + (1 - \alpha)m_3 \succ m_2 \succ \beta m_1 + (1 - \beta)m_3; \quad (7.50)$$

3. Independence: for all  $m_1$  and  $m_2$  and for all  $\alpha$  in  $(0, 1)$ ,  $m_1 \succ m_2$  implies

$$\alpha m_1 + (1 - \alpha)m_3 \succ \alpha m_2 + (1 - \alpha)m_3. \quad (7.51)$$

It is clear that  $U(m)$  in (7.49) becomes the expected utility when  $m$  is Bayesian: we then have

$$U(m) = \sum_{c \in \mathcal{C}} m(\{c\})u(c), \quad (7.52)$$

with  $u(c) = U(m_{\{c\}})$ . The major difference with the classical EU model is that we now need, in the general case, to elicit the utility values  $U(A)$  for each subset  $A \subseteq \mathcal{C}$  of consequences, which limits the practical use of the method. However, Jaffray [44] showed that a major simplification of (7.49) can be achieved by introducing an additional axiom. To present this axiom, let us introduce the following notation. Let us write  $c_1 \succcurlyeq c_2$  whenever  $m_{\{c_1\}} \succcurlyeq m_{\{c_2\}}$ . Furthermore, let  $\underline{c}_A$  and  $\bar{c}_A$  denote, respectively, the worst and the best consequence in  $A$ . The additional axiom can then be stated as follows:

4. Dominance: for all non-empty subsets  $A$  and  $B$  of  $\mathcal{C}$ , if  $\underline{c}_A \succcurlyeq \underline{c}_B$  and  $\bar{c}_A \succcurlyeq \bar{c}_B$ , then  $m_A \succcurlyeq m_B$ .

This axiom was justified by Jaffray [44] as follows. If  $\underline{c}_A \succcurlyeq \underline{c}_B$  and  $\bar{c}_A \succcurlyeq \bar{c}_B$ , it is possible to construct a set  $\Omega$  of states of nature, and two acts  $f : \Omega \rightarrow A$  and  $f' : \Omega \rightarrow B$ , such that, for any  $\omega \in \Omega$ ,  $f(\omega) \succcurlyeq f'(\omega)$ . As act  $f$  dominates  $f'$ , it should be preferred whatever the information on  $\Omega$ . Hence,  $f$  should be preferred to  $f'$  when we have a vacuous mass function on  $\Omega$ , in which case  $f$  and  $f'$  induce, respectively, the logical mass function  $m_A$  and  $m_B$  on  $\mathcal{C}$ .

The Dominance axiom immediately implies that, for any non-empty subsets  $A$  and  $B$  of  $\mathcal{C}$ , if  $\underline{c}_A \sim \underline{c}_B$  and  $\bar{c}_A \sim \bar{c}_B$ , then  $m_A \sim m_B$ , and  $U(A) = U(B)$ . Hence,  $U(A)$  depends only on the worst and the least consequence in  $A$ , and we can write  $U(A) = u(\underline{c}_A, \bar{c}_A)$ . Equation (7.49) thus becomes

$$U(m) = \sum_{\emptyset \neq A \subseteq \mathcal{C}} m(A)u(\underline{c}_A, \bar{c}_A). \quad (7.53)$$

If one accepts the Dominance axiom, one is led to rejecting the pignistic criterion, as well as the OWA criterion, except when it is identical to the Hurwicz criterion.

To describe the DM's attitude to ambiguity, Jaffray [44] then introduces the *local pessimism index*  $\alpha(\underline{c}, \bar{c})$ , defined as the value of  $\alpha$  which makes the DM indifferent between:

1. Receiving at least  $\underline{c}$  and at most  $\bar{c}$ , with no further information, and



2. Receiving either  $\underline{c}$  with probability  $\alpha$  or  $\bar{c}$  with probability  $1 - \alpha$ .

We then have

$$u(\underline{c}, \bar{c}) = \alpha(\underline{c}, \bar{c})u(\underline{c}) + (1 - \alpha(\underline{c}, \bar{c}))u(\bar{c}). \quad (7.54)$$

This relation shows how the DM's attitudes to risk and to ambiguity jointly determine  $U$ . Now, the utility of credibilistic lottery  $m$  can be written as

$$U(m) = \sum_{\emptyset \neq A \subseteq \mathcal{C}} m(A) [\alpha(\underline{c}_A, \bar{c}_A)u(\underline{c}_A) + (1 - \alpha(\underline{c}_A, \bar{c}_A))u(\bar{c}_A)]. \quad (7.55)$$

The Hurwicz criterion (7.40) corresponds to the case where  $\alpha(\underline{c}, \bar{c})$  is equal to a constant  $\alpha$ .

Jaffray's axioms are the counterpart of the axioms of Von Neumann and Morgenstern (see Section 7.2.2) for decision under risk: assuming uncertainty on the states of nature to be quantified by belief functions, they justify a particular decision strategy that maximizes the utility (7.55) of the credibilistic lotteries induced by the acts contemplated by the DM. Giang and Shenoy [37] have proposed an alternative set of axioms, leading to a different decision criterion (see also [36]). Their approach, however, is restricted to the case of partially consonant mass functions. A mass function  $m$  is said to be partially consonant if its focal sets can be divided into groups such that (a) the focal sets of different groups do not intersect and (b) the focal sets of the same group are nested. The family of partially consonant mass functions include Bayesian and consonant mass functions as special cases. In the context of statistical inference, Walley [104] has shown that partially consonant mass functions arise as a consequence of some axioms. There does not seem, however, to be any compelling reason to constrain belief functions to be partially consonant outside the specific context of statistical inference.

Until now, there is no decision-theoretic justification of belief functions similar to Savage's axioms for probabilities (see Section 7.2.3). There are, however, two remarkable results that justify the use of belief functions for decision-making. Interestingly, both results also justify Jaffray's decision criterion (7.55). The first of these results is due to Jaffray and Wakker [45], who consider the situation where objective probabilities are defined on a finite set  $S$ , and there is a multi-valued mapping  $\Gamma$  that maps each element  $s \in S$  to a subset  $\Gamma(s)$  of the set  $\Omega$  of states of nature. The authors thus do not postulate a belief function in the first place, but start one step before, i.e., they postulate the existence of a source, which induces a belief function  $Bel$  (see Section 2.1.2). Each act  $f : \Omega \rightarrow \mathcal{C}$  then carries  $Bel$  from  $\Omega$  to  $\mathcal{C}$ , i.e., it induces a credibilistic lottery. The authors then justify a neutrality axiom,

which states that two acts are indifferent whenever they generate the same credibilistic lottery. Finally, they derive the decision criterion (7.55) from two axioms: a continuity condition, and a weakened version of Savage's STP (7.17), which they call the *weak sure-thing principle* (WSTP). The WSTP is based on the notion of ambiguous event. A subset  $A \subseteq \Omega$  is said to be an ambiguous event if there is a focal set of  $Bel$  that intersects both  $A$  and  $\bar{A}$ ; other, it is an unambiguous event. The WSTP is satisfied if, for any two acts that have common outcomes outside an unambiguous event  $A \subset \Omega$ , the preference does not depend on the level of those common outcomes. Formally, for any unambiguous event  $A \subset \Omega$  and any acts  $f, g, h, h'$ , it is required that

$$fAh \succsim gAh \Rightarrow fAh' \succsim gAh'. \quad (7.56)$$

This weaker version of the STP, together with a technical continuity assumption, imply that preference among acts are represented by a utility function of the form (7.49). Criterion (7.55) is obtained by adding the Dominance axiom.

Interestingly, a similar criterion was obtained in [35], for the case where an act is defined as a multi-valued mapping from  $\Omega$  to  $2^{\mathcal{C}}$ . By postulating axioms similar to those of Savage, and two additional axioms, Ghirardato proved that the preference relation among acts is represented by a utility function similar to (7.55). The most important axiom in Ghirardato's construction, apart from the ones already considered by Savage, is the "contingencywise dominance" axiom, which states that, if a crisp act (i.e., an act such that  $f(\omega)$  is a singleton for any  $\omega \in \Omega$ ) is strictly better (resp., strictly worse) than all the possible elements of a set of consequences, then it should be preferred to receiving the set for sure (resp., vice versa).

## Appendix: the Choquet integral

Given the importance of the Choquet integral in modern decision analysis, we summarize some elements on this topic in this section. We also prove that the lower and upper expectations are Choquet integrals with respect to the belief and plausibility functions, respectively.

**Definition**

Let  $X$  be an integrable real random variable defined on a probability space  $(\Omega, \mathcal{A}, P)$ . Its expectation can be written as

$$\mathbb{E}_P(X) = \int_0^{+\infty} P(X \geq t) dt + \int_{-\infty}^0 [P(X \geq t) - 1] dt. \quad (7.57)$$

The Choquet integral [13] generalizes (7.57) by replacing the probability measure  $P$  with an arbitrary measure, defined as a function  $\mu : 2^\Omega \rightarrow \mathbb{R}$  such that  $\mu(\emptyset) = 0$ ,  $\mu(\Omega) = 1$  and, for any subsets  $E$  and  $F$  of  $\Omega$ ,

$$E \subseteq F \Rightarrow \mu(E) \leq \mu(F). \quad (7.58)$$

The Choquet integral of  $X : \Omega \rightarrow \mathbb{R}$  with respect to  $\mu$  is defined as

$$\mathbb{E}_\mu(X) = \int_0^{+\infty} \mu(X \geq t) dt + \int_{-\infty}^0 [\mu(X \geq t) - 1] dt. \quad (7.59)$$

Let us consider the case where  $\Omega = \{\omega_1, \dots, \omega_n\}$  is finite, and assume that the elements of  $\Omega$  have been indexed in such a way that

$$X(\omega_1) \leq X(\omega_2) \leq \dots \leq X(\omega_n). \quad (7.60)$$

Let  $A_k = \{\omega_k, \dots, \omega_n\}$  for  $1 \leq k \leq n$  and  $A_{n+1} = \emptyset$ . It can easily be checked that the Choquet integral of  $X$  has the following simple expression:

$$\mathbb{E}_\mu(X) = \sum_{i=1}^n X(\omega_i) [\mu(A_i) - \mu(A_{i+1})]. \quad (7.61)$$

We can observe that the weights  $\mu(A_i) - \mu(A_{i+1})$ , in the above expression, sum to one:  $\mathbb{E}_\mu(X)$  is thus a weighted average of the values taken by  $X$ , with weights depending on  $X$ .

**Properties and characterization**

The Choquet integral has the following properties:

1. Normalization:  $\mathbb{E}_\mu(\mathbb{1}_\Omega) = 1$ , where  $\mathbb{1}_\Omega(\cdot)$  is the indicator function of  $\Omega$ .
2. Comonotonic additivity: if  $f : \Omega \rightarrow \mathbb{R}$  and  $g : \Omega \rightarrow \mathbb{R}$  are *comonotonic*, i.e., if for every  $\omega, \omega'$  in  $\Omega$ , it holds that  $(f(\omega) - f(\omega'))(g(\omega) - g(\omega')) \geq 0$ , then

$$\mathbb{E}_\mu(f + g) = \mathbb{E}_\mu(f) + \mathbb{E}_\mu(g). \quad (7.62)$$

3. Monotonicity: for all  $f$  and  $g$ ,

$$f \geq g \Rightarrow \mathbb{E}_\mu(f) \geq \mathbb{E}_\mu(g). \quad (7.63)$$

Conversely, it can be shown [75] that, if a functional  $H$  from the space  $B$  of bounded real-valued functions on  $\Omega$  verifies the above three axioms, then it is the Choquet integral with respect to the measure  $\mu$  defined by  $\mu(A) = H(\mathbb{1}_A)$ , for all  $A \in 2^\Omega$ .

### Case of belief and plausibility measures

Let us now assume that the measure  $\mu$  is a belief function  $Bel$  induced by a mass function  $m$  on a finite frame  $\Omega = \{\omega_1, \dots, \omega_n\}$ . With the same notations as in (7.61), we have

$$\mathbb{E}_{Bel}(X) = \sum_{i=1}^n X(\omega_i) [Bel(A_i) - Bel(A_{i+1})]. \quad (7.64)$$

Now,

$$Bel(A_i) - Bel(A_{i+1}) = \sum_{B \subseteq A_i} m(B) - \sum_{B \subseteq A_{i+1}} m(B) \quad (7.65a)$$

$$= \sum_{B \subseteq A_i, \omega_i \in B} m(B). \quad (7.65b)$$

Hence,

$$\mathbb{E}_{Bel}(X) = \sum_{i=1}^n X(\omega_i) \sum_{B \subseteq A_i, \omega_i \in B} m(B). \quad (7.66)$$

For each subset  $B$  of  $\Omega$ ,  $m(B)$  is thus multiplied by  $X(\omega_i)$  for  $i$  such that  $X(\omega_i) = \min_{\omega \in B} X(\omega)$ . Consequently,

$$\mathbb{E}_{Bel}(X) = \sum_{B \subseteq \Omega} m(B) \min_{\omega \in B} X(\omega) = \underline{\mathbb{E}}(X). \quad (7.67)$$

The Choquet integral of  $X$  w.r.t.  $Bel$  is thus equal to the lower expectation of  $X$ .

Let us now consider the Choquet expectation of  $X$  w.r.t. the plausibility function  $Pl$ . We have

$$\mathbb{E}_{Pl}(X) = \sum_{i=1}^n X(\omega_i) [Pl(A_i) - Pl(A_{i+1})], \quad (7.68)$$

with

$$Pl(A_i) - Pl(A_{i+1}) = \sum_{B \cap A_i \neq \emptyset} m(B) - \sum_{B \cap A_{i+1} \neq \emptyset} m(B) \quad (7.69a)$$

$$= \sum_{B \cap A_i \neq \emptyset, B \cap A_{i+1} = \emptyset} m(B) \quad (7.69b)$$

$$= \sum_{B \cap A_{i+1} = \emptyset, \omega_i \in B} m(B). \quad (7.69c)$$

Now,  $B \cap A_{i+1} = \emptyset$  means that  $B$  does not contain any  $\omega_j$  such that  $j > i$ . Consequently,  $X(\omega_i)$  is the maximum of  $X(\omega)$  for  $\omega \in B$  and (7.68) can be written as

$$\mathbb{E}_{Pl}(X) = \sum_{B \subseteq \Omega} m(B) \max_{\omega \in B} X(\omega) = \bar{\mathbb{E}}(X). \quad (7.70)$$



## Chapter 8

# Belief Functions on Infinite Spaces

Until now, the presentation of belief functions has been restricted to the case where  $\Omega$  is finite. The theory of belief functions on finite frames is sufficient to represent expert opinions, because any infinite frame can always be coarsened to a finite one, which is more easily conceived by an expert. However, in some applications, the restriction to finite frames does appear as a limitation. For instance, in most statistical models, the parameter space is  $\mathbb{R}^d$  for some  $d \geq 1$ . It is thus useful to extend the theory from finite to infinite (continuous) spaces. This extension involves, in the most general case, considerably more mathematical sophistication than involved in the finite case [76, 78]. In the presentation below, we will try to avoid entering technical details and we will focus on the simplest models, which are sufficient for most applications, in particular for uncertainty propagation in numerical models, or for statistical inference.

In Chapter 2, we have noticed the formal connection between belief functions and random sets. This connection remains valid in the infinite case and, as the theory of random sets is well developed mathematically [67], it will provide a solid foundation for a theory of belief functions in infinite spaces.

### 8.1 General definitions and results

In the finite case, we derived the notion of belief function from that of mass function, and we later showed the equivalence with the complete monotonicity condition. In the infinite case, there may not be a mass function associated with a completely monotone function, so that we have to define a belief

function axiomatically from its properties (the most important one being complete monotonicity).

### 8.1.1 Definitions

Let us consider a set  $\Omega$  and an algebra  $\mathcal{B}$  of subsets of  $\Omega$  (see Section 1.3.1). A belief function on  $\mathcal{B}$  is a function  $Bel : \mathcal{B} \rightarrow [0, 1]$  verifying the following three conditions:

1.  $Bel(\emptyset) = 0$ ;
2.  $Bel(\Omega) = 1$ ;
3. For any  $k \geq 2$  and any collection  $B_1, \dots, B_k$  of elements of  $\mathcal{B}$ ,

$$Bel\left(\bigcup_{i=1}^k B_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} B_i\right). \quad (8.1)$$

Similarly, a plausibility function can be defined as a function  $Pl : \mathcal{B} \rightarrow [0, 1]$  such that:

1.  $Pl(\emptyset) = 0$ ;
2.  $Pl(\Omega) = 1$ ;
3. For any  $k \geq 2$  and any collection  $B_1, \dots, B_k$  of elements of  $\mathcal{B}$ ,

$$Pl\left(\bigcap_{i=1}^k B_i\right) \leq \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} Pl\left(\bigcup_{i \in I} B_i\right). \quad (8.2)$$

It is clear that, whenever  $Bel$  is a belief function,  $Pl$  defined by  $Pl(A) = 1 - Bel(\overline{A})$  is a plausibility function.

### 8.1.2 Belief function induced by a source

A convenient way to create a belief function is to define a source, i.e., a multivalued mapping from a probability space to  $\mathcal{B}$  [18]. More precisely, let  $S$  be a set,  $\mathcal{A}$  an algebra of subsets of  $S$ ,  $\mu$  a finitely additive probability measure on  $\mathcal{A}$ , and  $\Gamma : S \rightarrow 2^\Omega$  a multi-valued mapping. We can define two inverses of  $\Gamma$ :



1. The lower inverse

$$\Gamma_*(B) = B_* = \{s \in S \mid \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\}; \quad (8.3)$$

2. The upper inverse

$$\Gamma^*(B) = B^* = \{s \in S \mid \Gamma(s) \cap B \neq \emptyset\}, \quad (8.4)$$

for all  $B \in \mathcal{B}$ . We say that  $\Gamma$  is *strongly measurable* with respect to  $\mathcal{A}$  and  $\mathcal{B}$  iff, for all  $B \in \mathcal{B}$ ,  $B^* \in \mathcal{A}$ . This implies that, for all  $B \in \mathcal{B}$ ,  $B_* \in \mathcal{A}$ . To see this, we can notice that

$$B_* = \{s \in S \mid \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\} \quad (8.5a)$$

$$= \{s \in S \mid \Gamma(s) \subseteq B\} \cap \{s \in S \mid \Gamma(s) \neq \emptyset\} \quad (8.5b)$$

$$= \{s \in S \mid \Gamma(s) \cap \overline{B} = \emptyset\} \cap \{s \in S \mid \Gamma(s) \neq \emptyset\} \quad (8.5c)$$

$$= \overline{(\overline{B})^*} \cap \Omega^*. \quad (8.5d)$$

We then have the following important theorem [66].

**Theorem 8.1** *Let  $\mathcal{A}$  be an algebra of subsets of a set  $S$ ,  $\mu$  a finitely additive probability measure on  $\mathcal{A}$ ,  $\mathcal{B}$  an algebra of subsets of a set  $\Omega$ , and  $\Gamma$  a strongly measurable mapping w.r.t.  $\mathcal{A}$  and  $\mathcal{B}$  such that  $\mu(\Omega^*) \neq 0$ . Let the lower and upper probability measures on  $\mathcal{B}$  be defined as follows: for all  $B \in \mathcal{B}$ ,*

$$\mu_*(B) = K\mu(B_*), \quad (8.6a)$$

$$\mu^*(B) = K\mu(B^*) = 1 - \mu_*(\overline{B}), \quad (8.6b)$$

where  $K = [\mu(\Omega^*)]^{-1}$ . Then,  $\mu_*$  is a belief function and  $\mu^*$  is the dual plausibility function.

*Proof.* First, we have

$$\Omega^* = \{s \in S \mid \Gamma(s) \neq \emptyset\} \quad (8.7a)$$

$$= \{s \in S \mid \Gamma(s) \cap B \neq \emptyset\} \cup \{s \in S \mid \Gamma(s) \cap B = \emptyset, \Gamma(s) \neq \emptyset\} \quad (8.7b)$$

$$= B^* \cup (\overline{B})_*. \quad (8.7c)$$

Since  $B^* \cap (\overline{B})_* = \emptyset$ , we can deduce that  $\mu^*(B) = 1 - \mu_*(\overline{B})$ . To prove that  $\mu_*$  verifies (2.5), we can remark that

$$\Gamma_* \left( \bigcap_i B_i \right) = \bigcap_i \Gamma_*(B_i) \quad (8.8)$$

and

$$\Gamma^* \left( \bigcup_i B_i \right) \supseteq \bigcup_i \Gamma_*(B_i). \quad (8.9)$$

Consequently, for any  $k$  and any collection  $B_1, \dots, B_k$  of elements of  $\mathcal{B}$ ,

$$\begin{aligned} \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} \mu_* \left( \bigcap_{i \in I} B_i \right) &= \\ \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} K\mu \left( \bigcap_{i \in I} \Gamma_*(B_i) \right) &= K\mu \left( \bigcup_{i \in I} \Gamma_*(B_i) \right) \\ &\leq K\mu \left[ \Gamma_* \left( \bigcup_i B_i \right) \right] = \mu_* \left( \bigcup_i B_i \right). \end{aligned} \quad (8.10)$$

Thus, to define a belief function on  $(\Omega, \mathcal{B})$ , it suffices to define a finitely additive probability  $\mu$  on an algebra  $\mathcal{A}$  of subsets of a set  $S$  and a strongly measurable mapping  $\Gamma$  from  $S$  to  $\mathcal{A}$ . By analogy with the finite case, the sets  $\Gamma(s)$  for  $s \in S$  can be called the focal sets of  $Bel$ .

Conversely, Shafer [76] showed that any belief function  $Bel$  on an algebra  $\mathcal{B}$  is induced by a source  $(S, \mathcal{A}, \mu, \Gamma)$ , where  $\mathcal{A}$  is an algebra and  $\mu$  a finitely additive probability. In Shafer's constructive proof (too complex to be reproduced here),  $S = 2^\Omega$ ,  $\mathcal{A}$  is an algebra of subsets of  $2^\Omega$ , and  $\Gamma$  is the identity mapping. However, as well shall see in Section 8.2, a much simpler representation can often be found in practice.

As shown by Shafer [78], any belief function  $Bel$  on  $(\Omega, \mathcal{B})$  can be extended to  $(\Omega, 2^\Omega)$  as

$$\widetilde{Bel}(A) = \sup\{Bel(B) \mid B \in \mathcal{B}, B \subseteq A\}, \quad (8.11)$$

for all  $A \subseteq \Omega$ . A belief function  $Bel$  on a  $\sigma$ -algebra  $\mathcal{B}$  is *continuous* if, for any decreasing sequence  $B_1 \supset B_2 \supset B_3 \supset \dots$  of elements of  $\mathcal{B}$ ,

$$\lim_{i \rightarrow +\infty} Bel(B_i) = Bel \left( \bigcap_i B_i \right). \quad (8.12)$$

If the algebras  $\mathcal{A}$  and  $\mathcal{B}$  of Theorem 8.1 are  $\sigma$ -algebras, and if  $\mu$  is a countably additive probability, then the belief function  $Bel$  on  $\mathcal{B}$  induced by a strongly measurable mapping  $\Gamma$  is continuous. To see this, we can notice that, for any decreasing sequence  $B_1 \supset B_2 \supset B_3 \supset \dots$  of elements of  $\mathcal{B}$ ,

$\Gamma_*(B_1) \supset \Gamma_*(B_2) \supset \Gamma_*(B_3) \supset \dots$  is a decreasing sequence of elements of  $\mathcal{A}$ . Consequently,

$$\begin{aligned} \lim_{i \rightarrow +\infty} \mu_*(B_i) &= \lim_{i \rightarrow +\infty} K\mu(\Gamma_*(B_i)) = \\ K\mu\left(\bigcap_i \Gamma_*(B_i)\right) &= K\mu\left[\Gamma_*\left(\bigcap_i B_i\right)\right] = \mu_*\left(\bigcap_i B_i\right). \end{aligned} \quad (8.13)$$

### 8.1.3 Dempster's rule

Assume that we have  $n$  sources  $(S_i, \mathcal{A}_i, \mu_i, \Gamma_i)$  for  $i = 1, \dots, n$ , where each  $\Gamma_i$  is a multi-valued mapping from  $S_i$  to  $2^\Omega$ . Then, the combined source  $(S, \mathcal{A}, \mu, \Gamma_\cap)$  can be defined as follows [18]:

$$S = S_1 \times S_2 \dots \times S_n, \quad (8.14a)$$

$$\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2 \dots \otimes \mathcal{A}_n, \quad (8.14b)$$

$$\mu = \mu_1 \times \mu_2 \dots \times \mu_n, \quad (8.14c)$$

$$\Gamma_\cap(s) = \Gamma_1(s_1) \cap \Gamma_2(s_2) \cap \dots \cap \Gamma_n(s_n), \quad (8.14d)$$

where  $\mathcal{A}$  is the algebra generated by subsets of  $S$  of the form  $A_1 \times \dots \times A_n$ , with  $A_i \in \mathcal{A}_i$ , and  $\mu$  is the product measure. The belief function  $Bel$  induced by  $(S, \mathcal{A}, \mu, \Gamma_\cap)$  can then be written as  $Bel_1 \oplus \dots \oplus Bel_n$ , where  $Bel_i$  is the belief function induced by source  $i$ . For each  $B \in \mathcal{B}$ ,  $Bel(B)$  is the conditional probability that  $\Gamma_\cap(s) \subseteq B$ , given that  $\Gamma_\cap(s) \neq \emptyset$ ,

$$Bel(B) = \frac{\mu(\{s \in S | \Gamma_\cap(s) \neq \emptyset, \Gamma_\cap(s) \subseteq B\})}{\mu(\{s \in S | \Gamma_\cap(s) \neq \emptyset\})}, \quad (8.15)$$

which is well defined iff the denominator is non null (i.e., if the  $n$  belief functions are not totally conflicting). As in the finite case, the degree of conflict between the  $n$  belief functions can be defined as one minus the denominator in (8.15).

This combination rule, introduced by Dempster in Ref. [18], clearly generalizes that introduced in Chapter 3 for the finite case. The consideration of the product probability measure in (8.14c) corresponds to an assumption of independence between the items of evidence, as discussed in Chapter 3.

A source  $(S, \mathcal{A}, \mu, \Gamma)$  induces a commonality function  $Q$  if for any  $B \in \mathcal{B}$ , the set  $\{s \in S | \Gamma(s) \supseteq B\}$  is in  $\mathcal{A}_i$ . We can then define the commonality function as

$$Q(B) = K\mu(\{s \in S | \Gamma(s) \supseteq B\}), \quad (8.16)$$

for all  $B \in \mathcal{B}$ , with  $K = [\mu(\Omega^*)]^{-1}$ . If each of the  $n$  sources  $(S_i, \mathcal{A}_i, \mu_i, \Gamma_i)$  induces a commonality function  $Q_i$ , then so does the combined source  $(S, \mathcal{A}, \mu, \Gamma_\cap)$ , and we can easily check that

$$Q(B) = \frac{1}{1 - \kappa} \prod_{i=1}^n Q_i. \quad (8.17)$$

## 8.2 Practical models

In this section, we describe two special cases of belief functions on  $\mathbb{R}^d$  that will play an important role in applications, especially for statistical inference: consonant random closed sets and random intervals.

### 8.2.1 Consonant random closed sets

Let us assume that  $\Omega = \mathbb{R}^d$  and  $\mathcal{B} = 2^\Omega$ . Let  $\pi$  be an upper semi-continuous map from  $\mathbb{R}^d$  to  $[0, 1]$ , i.e., for any  $s \in [0, 1]$ , the set

$${}^s\pi = \{x \in \mathbb{R}^d \mid \pi(x) \geq s\} \quad (8.18)$$

is closed. Furthermore, assume that  $\pi(x) = 1$  for some  $x$ . Let  $S = [0, 1]$ ,  $\mathcal{A}$  be the Borel  $\sigma$ -field on  $[0, 1]$ ,  $\mu$  the Lebesgue measure, and  $\Gamma$  the mapping defined by  $\Gamma(s) = {}^s\pi$ . Then  $\Gamma$  is strongly measurable and it defines a random closed set [67]. We can observe that its focal sets are nested: it is said to be *consonant*. Let  $Bel$ ,  $Pl$  and  $Q$  be the corresponding belief, plausibility and commonality functions. Then, for any  $B \subseteq \mathbb{R}^d$ :

$$Pl(B) = \mu(\{s \in S \mid {}^s\pi \cap B \neq \emptyset\}) \quad (8.19a)$$

$$= \mu(\{s \in S \mid \exists x \in B, \pi(x) \geq s\}) \quad (8.19b)$$

$$= \mu(\{s \in S \mid s \leq \sup_{x \in B} \pi(x)\}) \quad (8.19c)$$

$$= \sup_{x \in B} \pi(x), \quad (8.19d)$$

$$Bel(B) = 1 - Pl(\overline{B}) = 1 - \sup_{x \notin B} \pi(x) = \inf_{x \notin B} (1 - \pi(x)) \quad (8.20)$$

and

$$Q(B) = \mu(\{s \in S \mid {}^s\pi \supseteq B\}) \quad (8.21a)$$

$$= \mu(\{s \in S \mid \forall x \in B, \pi(x) \geq s\}) \quad (8.21b)$$

$$= \mu(\{s \in S \mid s \leq \inf_{x \in B} \pi(x)\}) \quad (8.21c)$$

$$= \inf_{x \in B} \pi(x). \quad (8.21d)$$

In particular,  $Pl\{x\} = Q(\{x\}) = \pi(x)$  for all  $x$ .

We will see an example of a belief function induced by a random closed set in Section 9.3.

### 8.2.2 Random closed intervals

In this section, we consider the case where  $\Omega = \mathbb{R}$ . In this case, a special class of random closed set is of special interest: random closed intervals [21].

#### Definition and properties

Let  $(U, V)$  be a bi-dimensional random vector from  $(S, \mathcal{A}, \mu)$  to  $\mathbb{R}^2$  such that

$$\mu(\{s \in S | U(s) \leq V(s)\}) = 1. \quad (8.22)$$

The mapping

$$\Gamma : s \rightarrow \Gamma(s) = [U(s), V(s)], \quad (8.23)$$

is strongly measurable. It defines a random closed interval. Two special cases are of interest:

1. If the random vector  $(U, V)$  is discrete, with  $\mu(U = u_i; V = V_i) = m_i$ , we have a discrete random interval; it is characterized by a mass function  $m$  with focal sets  $I_i = [u_i, v_i]$  and masses  $m(I_i) = m_i$ .
2. If  $(U, V)$  is absolutely continuous with density  $f(u, v)$ , we have a continuous random interval.

For all  $x \in \mathbb{R}$ , we have:

$$Bel((-\infty, x]) = \mu([U, V] \subseteq (-\infty, x]) = \mu(V \leq x) = F_V(x), \quad (8.24)$$

where  $F_V$  is the cumulative distribution function (cdf) of  $V$ , and

$$Pl((-\infty, x]) = \mu([U, V] \cap (-\infty, x] \neq \emptyset) = \mu(U \leq x) = F_U(x). \quad (8.25)$$

These functions are called, respectively, the lower and upper cdf of  $[U, V]$ . Now, for any  $a \leq b$ , we have

$$Bel([a, b]) = \mu([U, V] \subseteq [a, b]) = \mu(U \geq a; V \leq b), \quad (8.26)$$

$$\begin{aligned} Pl([a, b]) &= \mu([U, V] \cap [a, b] \neq \emptyset) = \\ &= 1 - \mu([U, V] \cap [a, b] = \emptyset) = 1 - \mu(V < a) - \mu(U > b) \end{aligned} \quad (8.27)$$

and

$$Q([a, b]) = \mu([U, V] \supseteq [a, b]) = \mu(U \leq a; V \geq b). \quad (8.28)$$

We can observe that if  $[U, V]$  is continuous, these probabilities can be computed by integrating the joint density  $f(u, v)$ . For instance,

$$Q([a, b]) = \int_{-\infty}^a \int_b^{+\infty} f(u, v) dv du, \quad (8.29)$$

$$Bel([a, b]) = \int_a^b \int_u^b f(u, v) dv du. \quad (8.30)$$

Conversely,

$$f(a, b) = -\frac{\partial^2 Q([a, b])}{\partial a \partial b} = -\frac{\partial^2 Bel([a, b])}{\partial a \partial b}. \quad (8.31)$$

### Lower and upper quantiles

If the random vector  $(U, V)$  is continuous, we can define its lower and upper quantiles at level  $\alpha$ , for any  $\alpha \in (0, 1)$ , as:

$$q_*(\alpha) = F_U^{-1}(\alpha), \quad (8.32a)$$

$$q^*(\alpha) = F_V^{-1}(\alpha). \quad (8.32b)$$

By definition,  $q_*(\alpha)$  and  $q^*(\alpha)$  are thus, respectively, the values such that

$$Pl((-\infty, q_*(\alpha)]) = \alpha \quad (8.33a)$$

and

$$Bel((-\infty, q^*(\alpha)]) = \alpha \quad (8.33b)$$

or, equivalently,

$$Pl((q^*(\alpha), +\infty)) = 1 - \alpha. \quad (8.33c)$$

For any  $\alpha \in (0, 0.5]$ , we may compute the  $\alpha$ -quantile interval  $(q_*(\alpha), q^*(1 - \alpha)]$ , which is such that  $Pl((-\infty, q_*(\alpha)]) = Pl((q^*(1 - \alpha), +\infty)) = \alpha$ . Because of the sub-additivity of  $Pl_x^Y$ , we may conclude that

$$Pl(\overline{(q_*(\alpha), q^*(1 - \alpha))}) \leq 2\alpha. \quad (8.34a)$$

or, equivalently,

$$Bel((q_*(\alpha), q^*(1 - \alpha))) \geq 1 - 2\alpha. \quad (8.34b)$$

The definitions of lower and upper quantiles can be extended to the case where  $(U, V)$  is discrete by linearly interpolating the lower and upper cdfs between the discrete values of  $U$  and  $V$ .

**Combination by Dempster's rule**

As in the finite case, random closed intervals can be combined using Dempster's rule. Let  $[U_1, V_1]$  and  $[U_2, V_2]$  be two random closed intervals, and let  $Q_1$  and  $Q_2$  be their commonality functions. We have the following equality:

$$(Q_1 \oplus Q_2)([a, b]) = \frac{1}{1 - \kappa} Q_1([a, b]) Q_2([a, b]), \quad (8.35)$$

where

$$\kappa = \mu([U_1, V_2] \cap [U_2, V_2] = \emptyset) \quad (8.36)$$

is the degree of conflict between the two random sets. To see this, we may observe that

$$(Q_1 \oplus Q_2)([a, b]) = \mu([U_1, V_1] \cap [U_2, V_2] \supseteq [a, b] | [U_1, V_1] \cap [U_2, V_2] \neq \emptyset), \quad (8.37)$$

which may be computed as

$$(Q_1 \oplus Q_2)([a, b]) = \frac{\mu([U_1, V_1] \supseteq [a, b], [U_2, V_2] \supseteq [a, b])}{\mu([U_1, V_1] \cap [U_2, V_2] \neq \emptyset)} \quad (8.38a)$$

$$= \frac{Q_1([a, b]) Q_2([a, b])}{1 - \kappa}. \quad (8.38b)$$

When  $[U_1, V_1]$  and  $[U_2, V_2]$  are continuous, the combination of  $[U_1, V_1] \oplus [U_2, V_2]$  may be cumbersome or even intractable. We may then compute an approximation, either by discretizing the two random intervals, or by using Monte Carlo simulation [5]. For instance, the following algorithm can be used to approximate  $(Pl_1 \oplus Pl_2)(A)$  for some  $A \subseteq \mathbb{R}$ :

```

k = 0
for i = 1 : N do
  Generate realizations  $[u_1, v_1]$  and  $[u_2, v_2]$  of  $[U_1, V_1]$  and  $[U_2, V_2]$ 
   $I = [u_1, v_1] \cap [u_2, v_2]$ 
  if  $I \cap A \neq \emptyset$  then
    k = k + 1
  end if
end for
 $(\widehat{Pl_1 \oplus Pl_2})(A) = \frac{k}{N}$ 

```





## Part II

# Application to Machine Learning and Statistical Inference



## Chapter 9

# Statistical Inference

In this chapter, we will consider the problem of modeling *statistical evidence* in the belief function framework. This problem actually motivated the first developments in the theory of belief functions [17, 18, 20, 21]. Statistical evidence consists in the observation of data generated by some random process. Based on this evidence, we wish to quantify the uncertainty about the random generating process itself, or about future data to be generated by the same or a related process. The former problem is referred to as *estimation*, the latter as *prediction*.

In this chapter, we will adopt the following notation. The observed data will be supposed to be a realization  $x$  of a random variable  $X$  taking values in a sample space  $\mathcal{X}$  (usually,  $\mathcal{X} = \mathbb{R}^n$  or  $\{0, 1\}^n$ ). The distribution of  $X$  is assumed to be known up to a parameter  $\theta \in \Theta$  and  $f_\theta(x)$  will denote its probability mass or density function. The *estimation* problem is to construct a belief function on  $\Theta$ , while the *prediction* problem is to construct a belief function about some new data  $Y$  to be generated from a random process depending on  $\theta$ .

In this chapter, we will first motivate the introduction of belief function-based methods by pointing to some limitations of classical approaches to statistical inference (Section 9.1). We will then present the two main methods of inference in the belief function framework: Dempster's method (Section 9.2) and the likelihood-based method (Section 9.3). A general prediction method will be then exposed in Section 9.4.

## 9.1 Limitations of classical approaches

There is a huge literature on statistical inference and we will not attempt to summarize it here. The purpose of this section is simply to underline some limitations of the classical approaches to statistical inference, which have motivated the development of new approaches based on belief functions.

### 9.1.1 Frequentist approach

The mainstream approach to statistical inference is the so-called frequentist approach, which essentially relies on confidence intervals (or, more generally, confidence sets) and significance testing.

A confidence set at level  $1 - \alpha$  is a random set that contains the true value of the parameter  $100(1 - \alpha)\%$  of the time, when samples are repeatedly drawn from the population. It is thus a *pre-experimental* measure of evidence, in so far as the accuracy of the evidence (the confidence level) is determined before the data is actually collected. Consequently, the confidence level does not qualify as a measure of the strength of the evidence for a particular realization of the random sample. To see this, consider the following example taken from [10].

**Example 9.1** *Suppose  $X_1$  and  $X_2$  are iid with probability mass function*

$$\mathbb{P}_\theta(X_i = \theta - 1) = \mathbb{P}_\theta(X_i = \theta + 1) = \frac{1}{2}, \quad i = 1, 2, \quad (9.1)$$

where  $\theta \in \mathbb{R}$  is an unknown parameter. Consider the following confidence set for  $\theta$ ,

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{otherwise.} \end{cases} \quad (9.2)$$

The corresponding confidence level  $1 - \alpha = P_\theta(\theta \in C(X_1, X_2))$  is

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_\theta(\theta \in C(X_1, X_2) | X_1 \neq X_2) \mathbb{P}_\theta(X_1 \neq X_2) + \\ &\quad \mathbb{P}_\theta(\theta \in C(X_1, X_2) | X_1 = X_2) \mathbb{P}_\theta(X_1 = X_2) = \\ &\quad 1 \times 0.5 + (0.5)^2 = 0.75. \end{aligned} \quad (9.3)$$

Now, let  $(x_1, x_2)$  be a given realization of the random sample  $(X_1, X_2)$ . If  $x_1 \neq x_2$ , we know for sure that  $\theta = (x_1 + x_2)/2$  and it would be absurd to take 75% as a measure of the strength of the statistical evidence. If  $x_1 = x_2$ , we know for sure that  $\theta$  is either  $x_1 - 1$  or  $x_1 + 1$ , but we have no reason

to favor any of these two hypotheses in particular. Again, it would make no sense to claim that the evidence support the hypothesis  $\theta = x_1 - 1$  with 75% confidence.  $\square$

The above example clearly shows that the confidence level is a pre-experimental measure that may not be relevant after seeing the data. From a more practical point of view, we may also notice that we often do not know how to construct exact confidence intervals for finite samples. Most confidence intervals used in practice are based, except for the simplest cases, on asymptotic assumptions. When used with small samples, these confidence intervals may have actual coverage probabilities quite different from their intended ones.

The other method for assessing the strength of statistical evidence in the frequentist approach is significance testing. Here, a hypothesis  $H \subset \Theta$  is contemplated, and a test statistic  $T$  is defined such that, say, large values of  $T$  are considered as evidence against  $H$ . Having observed a realization  $t$  of  $T$ , the  $p$ -value is defined as the probability of observing a value of  $T$  at least as large as  $t$ :  $p = P_H(T \geq t)$ . We can argue here that the logic of including not only the observed value  $t$ , but also more extreme values that have not been observed, is questionable. The problem with this approach is illustrated by the following example from [10].

**Example 9.2** Suppose the distributions of a discrete random variable  $X$  for two values  $\theta_1$  and  $\theta_2$  of some parameter  $\theta$  are given in the following table.

$x$	0	1	2	3	4
$f_{\theta_1}(x)$	0.75	0.14	0.04	0.037	0.033
$f_{\theta_2}(x)$	0.70	0.25	0.04	0.005	0.005

Let  $T = X$  be a test statistic for a significance test of either  $H_1 = \{\theta_1\}$  or  $H_2 = \{\theta_2\}$ , i.e., large values of  $X$  are considered as evidence against these hypotheses. If we have observed  $x = 2$ , the  $p$ -value against  $H_1$  is

$$p_1 = P_{\theta_1}(X \geq 2) = 0.04 + 0.037 + 0.033 = 0.11, \quad (9.4)$$

while the  $p$ -value against  $H_2$  is

$$p_2 = P_{\theta_2}(X \geq 2) = 0.04 + 0.005 + 0.005 = 0.05. \quad (9.5)$$

Thus, we would reject  $H_2$  at the 5% significance level, but we would not reject  $H_1$  at the same level. Yet, the probability of observed value,  $x = 2$ , is exactly the same under both hypotheses! There seems to be little ground for

including, in the calculation of the  $p$ -value, the probabilities of larger values of  $X$  that have not been observed.  $\square$

The fact that the  $p$ -value depends not only on observed data, but also on more extreme data that have not been observed, is a violation of the likelihood principle that will be exposed in Section 9.1.3.

To conclude this section, we may remark that frequentist methods are widely used for the interpretation of experimental data and they do yield sensible results in many situations. However, as shown by the example below, neither confidence sets nor significance tests provide post-experimental measures of the strength of statistical evidence.

### 9.1.2 Bayesian approach

The second main approach to statistical inference is the Bayesian approach, which treats the parameter as a random variable with prior distribution  $\pi(\theta)$ . The prior probabilities are usually considered as subjective and assumed to reflect the statistician's initial knowledge about the parameter, before observing the data. Considering the pdf  $f_\theta(x)$  as the conditional density  $f(x|\theta)$  of  $x$  given  $\theta$ , Bayes' rule allows us to compute the conditional pdf of  $\theta$  given the data  $x$ , referred to as the *posterior distribution* of  $\theta$ , and defined as

$$f(\theta|x) \propto \pi(\theta)f(x|\theta). \quad (9.6)$$

This method does not have the same problem as the confidence intervals and significance levels examined in the previous section, as the posterior distribution depends only on observed data, and is not based on averaging over the whole sample space. However, the critical issue is here the choice of the prior distribution in the (common) situation where we know nothing about  $\theta$  before observing the data. We should then select a "noninformative" prior distribution, but how can it be defined? If we can specify a bounded support for  $\theta$ , say, the interval  $[a, b]$ , then Laplace's Insufficient Reason Principle supports selecting a uniform distribution  $\mathcal{U}([a, b])$ . However, we have already stressed the main problem with this approach, which is that  $g(\theta)$  for a nonlinear function  $g$  will not have a uniform distribution, as illustrated by the wine-water paradox described in Section 1.3.5. Consequently, the Insufficient Reason Principle yields priors that are not invariant under reparameterization of the distribution of  $X$ .

Of course, this problem has been recognized for a long time and several remedies have been proposed. One of them is, in the case of a real parameter taking values in the real line, to define the prior distribution as  $\pi(\theta) = c$  for

all  $\theta \in \mathbb{R}$  and some constant  $c$ . Function  $\pi$  is, of course, no longer a pdf: it is said to be an *improper* distribution. This approach thus departs from “pure” probability theory, in which Bayesian inference is supposed to be grounded. Another attempt to define uninformative priors was made by Jeffreys [49]. The Jeffreys prior is defined objectively as being proportional to the square root of the determinant of the Fisher information.

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \quad (9.7)$$

where the component  $(i, j)$  of the information matrix  $I(\theta)_{ij}$  is

$$I(\theta)_{ij} = \mathbb{E}_{\theta} \left[ \frac{\partial \log f_{\theta}(x)}{\partial \theta_i} \frac{\partial \log f_{\theta}(x)}{\partial \theta_j} \right]. \quad (9.8)$$

The motivation for this definition is that the Jeffreys prior is invariant under reparameterization: if  $\varphi$  is a one-to-one transformation and  $\nu = \varphi(\theta)$ , then the Jeffreys prior on  $\nu$  is proportional to  $\sqrt{\det I(\nu)}$ . However, there are still some issues with this approach. First, the Jeffreys prior is sometimes improper. Secondly, and maybe more importantly, the Jeffreys prior can hardly be considered to be truly noninformative. For instance, consider an iid sample  $X_1, \dots, X_n$  from a Bernoulli distribution  $\mathcal{B}(\theta)$ . The Jeffreys prior on  $\theta$  is the beta distribution  $B(0.5, 0.5)$  whose pdf is displayed in Figure 9.1. We can see that extreme values of  $\theta$  are considered a priori more probable than central values, which does represent non vacuous knowledge about  $\theta$ .

### 9.1.3 Likelihood-based approach

Beside the frequentist and Bayesian schools of scientific inference, a third tradition can be traced back from Fisher’s later work [34] to Barnard [7], Birnbaum [12] and Edward [31], among others. The likelihood-based approach to statistical inference centers on direct inspection of the likelihood function  $L_x(\theta)$  alone, without relying on the concept of repeated sampling (underlying long-run frequency considerations) and without assuming the existence of a prior probability distribution. For proponents of this approach as Birnbaum [12], “reports of experimental results in scientific journals should in principle be descriptions of likelihood functions, when adequate mathematical models can be assumed, rather than reports of significance levels or interval estimates”.

The likelihood principle underlies the likelihood-based approach to statistical inference [12]. Let  $E$  denote a statistical model representing an experimental situation. Typically,  $E$  is composed of the parameter space  $\Theta$ ,

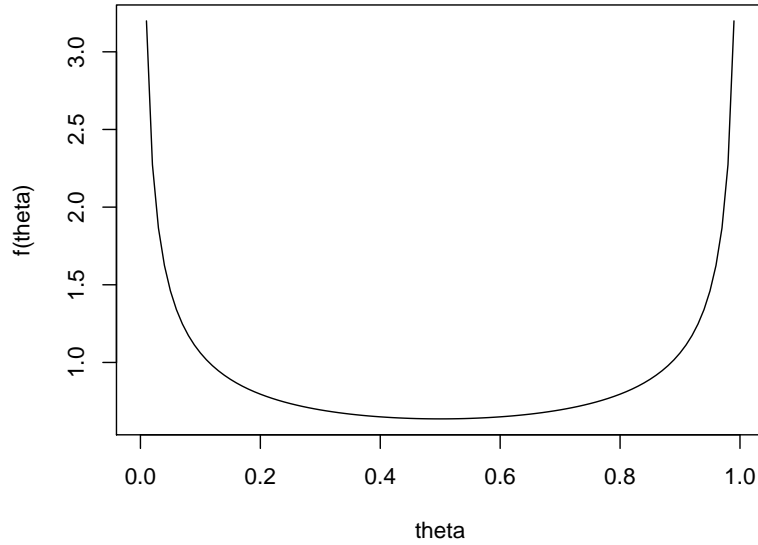


Figure 9.1: Jeffreys prior for a Bernoulli sample.

the sample space  $\mathcal{X}$  and a probability mass or density function  $f_\theta(x)$  for each  $\theta \in \Theta$ . Following Birnbaum [12], let us denote by  $Ev(E, x)$  the *evidential meaning* of the specified instance  $(E, x)$  of statistical evidence. The likelihood Principle (L) can be stated as follows:

*If  $E$  and  $E'$  are any two experiments with the same parameter space  $\Theta$ , represented by probability mass or density functions  $f_\theta(x)$  and  $g_\theta(y)$ , and if  $x$  and  $y$  are any two respective outcomes which determine likelihood functions satisfying  $f_\theta(x) = cg_\theta(y)$  for some positive constant  $c = c(x, y)$  and all  $\theta \in \Theta$ , then  $Ev(E, x) = Ev(E', y)$ .*

As noted by Birnbaum [12], the likelihood principle is an immediate consequence of Bayes' principle, which implies that the evidential meaning of  $(E, x)$  is contained in the posterior probability distribution  $p(\theta|x) \propto f_\theta(x)\pi(\theta)$ , where  $\pi(\theta)$  is the prior probability distribution. However, it was also accepted as self-evident by statisticians who did not adhere to the Bayesian school, including Fisher [34] and Barnard [7]. From a non Bayesian perspective, it was placed on firm ground by Birnbaum [12], who showed that it can be derived from the principles of sufficiency and conditionality, which are accepted by most (but not all) statisticians.

As already remarked above, Bayesian inference complies with the likelihood principle. However, for a Bayesian statistician, the likelihood function



alone does not constitute a valid representation of the statistical evidence, it needs to be multiplied by a prior. Frequentist methods, in contrast, do not comply with the likelihood principle. For instance, consider an urn with a proportion  $\theta$  of black balls, and the following two experiments:

- Experiment 1: a fixed number  $n$  of balls are drawn with replacement from the urn and the number  $X$  of black balls is observed;  $X$  has a binomial distribution  $\mathcal{B}(n, \theta)$ .
- Experiment 2: balls are drawn with replacement from the urn until a fixed number  $x$  of black balls have been drawn; we observe the number  $N$  of draws, which has a negative binomial distribution.

Confidence intervals computed in these two cases are different, although the likelihood functions for these two experiments are identical. This is because confidence intervals (and significance tests) depend not only on the likelihood, but also on the sample space.

The concept of likelihood function is clear from a statistical point of view, but it does not fit clearly in the more general landscape of uncertainty theories. Fisher, who introduced the likelihood function [31, 2], repeatedly stressed that “probability and likelihood are quantities of an entirely different nature” [33] as, in particular, likelihoods are not additive. In Section 9.3, we will present an approach that links the notion of likelihood function to that of belief function, allowing us to represent statistical evidence in the Dempster-Shafer framework. Before that, we will study in the next section the method of inference initially proposed by Dempster, based on different ideas.

## 9.2 Dempster’s method

The starting point of the theory of belief functions is a series of papers published by Dempster in 1960 [17, 18, 19, 20, 21], in which a new method of inference based on a multi-valued mapping was introduced. A recent account of this method can be found in Dempster’s later work [22]. Similar ideas are explored in [55, 56, 62] with a different terminology.

### 9.2.1 General method

The key idea underlying Dempster’s method of inference is to model the data-generating mechanism by an equation

$$X = \varphi(\theta, W) \tag{9.9}$$

that relates the data  $X$ , the parameter  $\theta$  and an unobserved auxiliary variable  $W$  with sample space  $\mathcal{W}$  and known probability distribution  $\mathbb{P}_W$  independent of  $\theta$ , such that, for any measurable subset  $A$  of  $\mathcal{X}$ ,

$$\mathbb{P}_\theta(X \in A) = \mathbb{P}_W(\varphi(\theta, W) \in A). \quad (9.10)$$

When  $X$  is a one-dimensional continuous random variable, a natural choice for  $W$  is  $W = F_\theta(X)$ , where  $F_\theta$  is the cumulative distribution function (cdf) of  $X$ . The random variable  $W$  then has then a uniform distribution on  $[0, 1]$  and (9.9) becomes

$$X = F_\theta^{-1}(W). \quad (9.11)$$

When  $W$  is discrete, (9.11) is still valid if  $F_\theta^{-1}$  now denotes the generalized inverse of  $F_\theta$ ,

$$F_\theta^{-1}(w) = \inf\{x | F_\theta(x) \geq w\}. \quad (9.12)$$

From now on, we will assume that  $\mathcal{W} = [0, 1]$  and  $\mathbb{P}_W$  is the uniform probability measure on  $[0, 1]$ .

Equation (9.9) defines a multi-valued mapping from  $[0, 1]$  to  $2^{\mathcal{X} \times \Theta}$  that maps each  $W$  to the pairs  $(X, \theta)$  compatible with  $W$ ,

$$\Gamma : W \rightarrow \Gamma(W) = \{(x, \theta) \in \mathcal{X} \times \Theta | x = \varphi(\theta, W)\}. \quad (9.13)$$

The four-tuple  $([0, 1], \mathcal{B}([0, 1]), \mathbb{P} - W, \Gamma)$ , where  $\mathcal{B}([0, 1])$  is the Borel sigma-field on  $[0, 1]$ , is a source for a belief function  $Bel^{\mathcal{X} \times \Theta}$  on  $\mathcal{X} \times \Theta$  (see Section 8.1.2).

Assume that  $\theta$  is known. Our belief about  $X$  is obtained by conditioning  $Bel^{\mathcal{X} \times \Theta}$  on  $\mathcal{X} \times \{\theta\}$  and marginalizing on  $\mathcal{X}$ . The multi-valued mapping (9.13) becomes

$$\Gamma_\theta : W \rightarrow \Gamma(W) = \{x \in \mathcal{X} | x = \varphi(\theta, W)\} = \{\varphi(\theta, W)\}. \quad (9.14)$$

By construction, for any measurable subset  $A$  of  $\mathcal{X}$ ,

$$Bel^{\mathcal{X}}(A) = \mathbb{P}_W(\varphi(\theta, W) \in A) = \mathbb{P}_\theta(X \in A), \quad (9.15)$$

i.e.,  $Bel^{\mathcal{X}}$  is the probability distribution of  $X$  given  $\theta$ .

Symmetrically, assume that we observe  $X = x$ . Our belief in  $\theta$  is obtained by conditioning  $Bel^{\mathcal{X} \times \Theta}$  on  $\{x\} \times \Theta$  and marginalizing on  $\Theta$ . The corresponding multi-valued mapping is

$$\Gamma_x : W \rightarrow \Gamma(W) = \{\theta \in \Theta | x = \varphi(\theta, W)\}. \quad (9.16)$$

The source  $([0, 1], \mathcal{B}([0, 1]), \mathbb{P}_W, \Gamma_x)$  defines a belief function  $Bel_x^\Theta$  on  $\Theta$ , which represents the statistical evidence.

This method has the following important property. Let  $X_1, \dots, X_n$  be an iid sample from  $f_\theta(x)$ . Each  $X_i$  can be written as  $X_i = \varphi(\theta, W_i)$ , where  $W_1, \dots, W_n$  are iid from the standard uniform distribution. The belief function on  $\Theta$  after observing  $\mathbf{x} = (x_1, \dots, x_n)$  is obtained by combining each  $Bel_{x_i}^\Theta$  by Dempster's rule,

$$Bel_{\mathbf{x}}^\Theta = Bel_{x_1}^\Theta \oplus Bel_{x_2}^\Theta \oplus \dots \oplus Bel_{x_n}^\Theta. \quad (9.17)$$

### 9.2.2 Application to a Bernoulli sample

Let us assume that we have an urn with an unknown proportion  $\theta$  of black balls, and consider the experiment that consists of drawing  $n$  balls with replacement from the urn. The result of this experiment can be denoted by  $X_1, \dots, X_n$ , where  $X_i = 1$  if the  $i$ -th ball drawn from the urn was black and  $X_i = 0$  otherwise. It is an iid sample from the Bernoulli  $\mathcal{B}(\theta)$  distribution. Based on these observations, what can be said about the composition of the urn? The solution of this problem using Dempster's method (given in [17]) is described in this section.

The  $\varphi$ -equation (9.9) in this case is

$$X_i = \varphi(\theta, W_i) = \begin{cases} 1 & \text{if } W_i \leq \theta \\ 0 & \text{otherwise,} \end{cases} \quad (9.18)$$

where  $W_i$  has a standard uniform distribution. The multi-valued mapping  $\Gamma$  from (9.14) is

$$\Gamma(W) = (\{0\} \times [0, W]) \cup (\{1\} \times [W, 1]). \quad (9.19)$$

The form of the corresponding focal sets is illustrated in Figure 9.2.

After conditioning on  $x$ , we get

$$\Gamma_x(W) = \begin{cases} [0, W] & \text{if } x = 0, \\ [W, 1] & \text{if } x = 1, \end{cases} \quad (9.20)$$

which defines a random interval (see Section 8.2.2). The corresponding commonality function can be computed as follows. If  $x = 0$ , for any  $0 \leq u \leq v \leq 1$ ,

$$Q_x([u, v]) = \mathbb{P}_W([0, W] \supseteq [u, v]) = \mathbb{P}_W(W > v) = 1 - v. \quad (9.21)$$

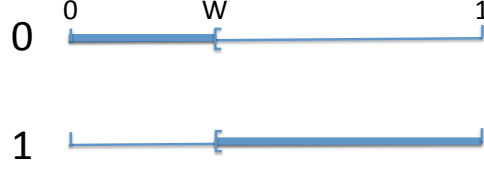


Figure 9.2: Focal sets  $(\{0\} \times [0, W]) \cup (\{1\} \times [W, 1])$  of the belief function  $Bel^{X \times \Theta}$  in the case of a Bernoulli sample.

If  $x = 1$ ,

$$Q_x([u, v]) = \mathbb{P}_W([w, 1] \supseteq [u, v]) = \mathbb{P}_W(W \leq u) = u. \quad (9.22)$$

After observing  $n$  realizations  $x_1, \dots, x_n$ , the commonality function becomes, from (9.17) and (8.35),

$$Q([u, v]) \propto \prod_{i=1}^n Q_{x_i}([u, v]) = u^N (1-v)^{n-N}, \quad (9.23)$$

where  $N = \sum_{i=1}^n x_i$ . The joint density  $f(u, v)$  of the bounds  $U$  and  $V$  of the random interval can be obtained from (8.31), by differentiating  $Q([u, v])$ . If  $0 < N < n$ ,

$$f(u, v) = -\frac{\partial^2 Q([u, v])}{\partial u \partial v} = c u^{N-1} (1-v)^{n-N-1}, \quad (9.24)$$

where the proportionality constant can be shown to be

$$c = N(n-N)C_n^N. \quad (9.25)$$

The cases where  $N = 0$  and  $N = n$  require special treatment. If  $N = 0$ , all the focal sets are of the form  $[0, v)$  and

$$Q([u, v]) = (1-v)^n. \quad (9.26)$$

(Note that the degree of conflict in Dempster's combination is equal to zero in this case). The lower bound  $U$  is then constant, while the upper bound has density

$$f(v) = -\frac{\partial^2 Q([0, v])}{\partial v} = (1-v)^{n-1}. \quad (9.27)$$

If  $N = n$ , the focal sets are of the form  $[u, 1]$  and

$$Q([u, v]) = u^n. \quad (9.28)$$

The lower bound  $V$  is constantly equal to 1, while the lower bound has density

$$f(u) = -\frac{\partial^2 Q([u, 1])}{\partial u} = nu^{n-1}. \quad (9.29)$$

We can remark that, if exactly  $N$  of the  $x_i$ 's equal 1, we can deduce that  $N$  of the  $w_i$ 's are less than or equal to  $\theta$ , and  $n - N$  are strictly greater than  $\theta$ . The intersection of the  $N$  intervals  $[0, w_i]$  for  $x_i = 0$  and the  $n - N$  intervals  $[w_i, 1]$  for  $x_i = 1$  is thus  $[w_{(N)}, w_{(N+1)})$ , where  $w_{(k)}$  is the  $k$ -th order statistics of the iid uniform sample  $W_1, \dots, W_n$ . Consequently,  $f(u, v)$  is the pdf of  $(W_{(N)}, W_{(N+1)})$ . The marginal distributions of  $U$  and  $V$  are, respectively, the Beta distributions  $B(N, n - N + 1)$  and  $B(N + 1, n - N)$ . The upper and lower cdf of  $\theta$  are thus easily computed from (8.24) and (8.25) as

$$Bel(\theta \leq t) = F_{B(N, n-N+1)}(t), \quad (9.30a)$$

$$Pl(\theta \leq t) = F_{B(N+1, n-N)}(t), \quad (9.30b)$$

where  $F_{B(\alpha, \beta)}$  denotes the cdf of the Beta distribution  $B(\alpha, \beta)$ . Given that the expectation of the distribution  $B(\alpha, \beta)$  is  $\alpha/(\alpha + \beta)$ , the lower and upper expectations of  $\theta$  are obtained from (??) and (??) as

$$\underline{\mathbb{E}}(\theta) = \mathbb{E}(U) = \frac{N}{n + 1}, \quad (9.31a)$$

$$\overline{\mathbb{E}}(\theta) = \mathbb{E}(V) = \frac{N + 1}{n + 1}. \quad (9.31b)$$

### 9.3 Likelihood-based method

Although based on simple principles, Dempster's method of inference quickly leads to intricate derivations, as shown in Section 9.2.2, where it was applied to one of the simplest statistical models. In this section, we will introduce another method in which a belief function in the parameter space is constructed from the likelihood function. This approach, first introduced by Shafer in [77], was later studied by Wasserman [106] and Aickin [1], among others. It was recently justified axiomatically by Denœux [25].

#### 9.3.1 General method

Given the statistical model  $f_\theta(x)$ ,  $x \in \mathcal{X}$ ,  $\theta \in \Theta$ , assume that we observe a realization  $x$  of the random variable  $X$ . We wish to represent the information gained on parameter  $\theta$  by a belief function  $Bel_x^\Theta$ . Which requirements should be imposed on  $Bel_x^\Theta$ ? The following three requirements seem relevant [25]:

1. Likelihood principle:  $Bel_x^\Theta$  should only depend on the likelihood function. As noted in Section 9.1.3, the likelihood principle has a solid foundation, as it can be derived from the two generally accepted principles of exhaustivity and conditionality [12].
2. Compatibility with Bayesian inference: if a Bayesian prior  $\pi(\theta)$  is available, combining it with  $Bel_x^\Theta$  using Dempster's rule should yield the Bayesian posterior.
3. Least Commitment Principle (see Chapter 4):  $Bel_x^\Theta$  should be the least committed belief function, among all those satisfying the previous two requirements.

The first two requirements jointly imply that the contour function  $pl_x(\theta)$  associated to  $Bel_x^\Theta$  should be proportional to the likelihood function:

$$pl_x(\theta) \propto L_x(\theta), \quad (9.32)$$

for all  $\theta \in \Theta$ . The least committed belief function, according to the commonality ordering  $\sqsubseteq_Q$  introduced in Section 4.1.1, among all those verifying (9.32), is the consonant belief function whose contour function is the relative likelihood function,

$$pl_x(\theta) = \frac{L_x(\theta)}{\sup_{\theta' \in \Theta} L_x(\theta')}. \quad (9.33)$$

for all  $\theta \in \Theta$ , where it is assumed that  $\sup_{\theta' \in \Theta} L_x(\theta') < \infty$ .

This belief function is called the *likelihood-based belief function* on  $\Theta$  induced by  $x$ . The corresponding plausibility function can be computed from  $pl_x$  as:

$$Pl_x^\Theta(A) = \sup_{\theta \in A} pl_x(\theta), \quad (9.34)$$

for all  $A \subseteq \Theta$ . The focal sets of  $Bel_x^\Theta$  are the levels sets of  $pl_x(\theta)$  defined as follows:

$$\Gamma_x(S) = \{\theta \in \Theta | pl_x(\theta) \geq S\}, \quad (9.35)$$

for  $S \in [0, 1]$ . These sets may be called *plausibility regions* and can be interpreted as sets of parameter values whose plausibility is greater than some threshold  $S$ . If  $S$  is considered to be random with continuous uniform distribution  $\mathbb{P}_S$  in  $[0, 1]$ , then the four-tuple  $([0, 1], \mathcal{B}([0, 1]), \mathbb{P}_S, \Gamma)$  is a source for the belief function  $Bel_x^\Theta$  (see Section 8.2.1). In particular, the following equalities hold:

$$Bel_x^\Theta(A) = \mathbb{P}_S(\{\Gamma_x(S) \subseteq A\}) \quad (9.36a)$$

$$Pl_x^\Theta(A) = \mathbb{P}_S(\{\Gamma_x(S) \cap A \neq \emptyset\}), \quad (9.36b)$$

for all  $A \subseteq \Theta$  such that the above expressions are well-defined.

We can also remark that the maximum likelihood estimate (MLE) of  $\theta$  can be interpreted as the value of  $\theta$  with the highest plausibility, and likelihood regions as defined by Edwards [31], among others, are identical to plausibility regions.

### 9.3.2 Bernoulli example

As an example, let us consider the same model as in Section 9.2.2. The likelihood function after observing a realization  $\mathbf{x} = (x_1, \dots, x_n)$  of the iid random sample  $X_1, \dots, X_n$  is

$$L_{\mathbf{x}}(\theta) = \theta^N (1 - \theta)^{n-N} \quad (9.37)$$

with  $N = \sum_{i=1}^n x_i$ . The likelihood-based belief function induced by  $x$  has the following contour function:

$$pl_{\mathbf{x}}(\theta) = \frac{\theta^x (1 - \theta)^{n-N}}{\widehat{\theta}^N (1 - \widehat{\theta})^{n-N}} = \left( \frac{\theta}{\widehat{\theta}} \right)^{n\widehat{\theta}} \left( \frac{1 - \theta}{1 - \widehat{\theta}} \right)^{n(1-\widehat{\theta})}, \quad (9.38)$$

for all  $\theta \in \Theta = [0, 1]$ , where  $\widehat{\theta} = N/n$  is the maximum likelihood estimate (MLE) of  $\theta$ . Function  $pl_{\mathbf{x}}(\theta)$  is plotted in Figure 9.3 for  $\widehat{\theta} = 0.4$  and  $n \in \{10, 20, 100\}$ . We can see that the contour function becomes more specific as  $n$  increases.

As  $pl_{\mathbf{x}}(\theta)$  is unimodal and continuous, each plausibility region  $\Gamma_x(S)$  for  $S \in [0, 1]$  is a closed interval  $[U(S), V(S)]$  and  $Bel_{\mathbf{x}}^{\Theta}$  is equivalent to the a closed random interval  $[U, V]$  [21]. The marginal cumulative probability distribution of  $U$  and  $V$  can be obtained as follows:

$$F_U(u) = \mathbb{P}(U \leq u) \quad (9.39a)$$

$$= \mathbb{P}([U, V] \cap (-\infty, u] \neq \emptyset) \quad (9.39b)$$

$$= Pl_{\mathbf{x}}^{\Theta}((-\infty, u]) \quad (9.39c)$$

$$= \begin{cases} pl_{\mathbf{x}}(u) & \text{if } u \leq \widehat{\theta} \\ 1 & \text{otherwise,} \end{cases} \quad (9.39d)$$

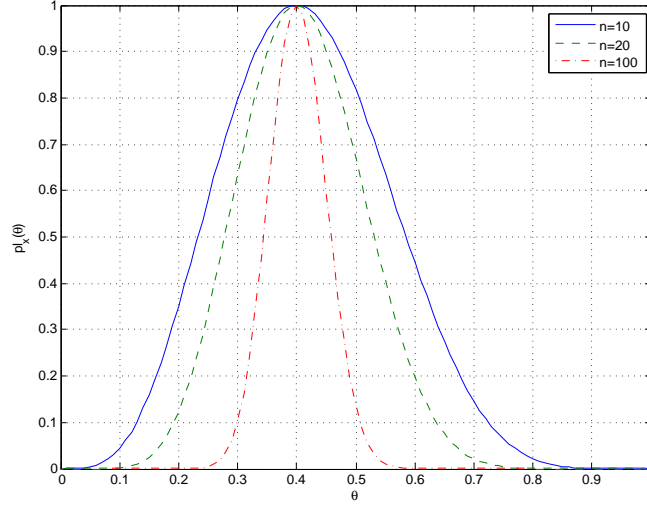


Figure 9.3: Contour functions (normalized likelihood functions) for the binomial distribution with  $\hat{\theta} = 0.4$  and  $n \in \{10, 20, 100\}$ .

and

$$F_V(v) = \mathbb{P}(V \leq v) \quad (9.40a)$$

$$= \mathbb{P}([U, V] \subseteq (-\infty, v]) \quad (9.40b)$$

$$= Bel_{\mathbf{x}}^{\Theta}((-\infty, v]) \quad (9.40c)$$

$$= 1 - Pl_{\mathbf{x}}^{\Theta}((v, +\infty)) \quad (9.40d)$$

$$= \begin{cases} 0 & \text{if } v \leq \hat{\theta} \\ 1 - pl_{\mathbf{x}}(v) & \text{otherwise.} \end{cases} \quad (9.40e)$$

### 9.3.3 Properties

Viewing the relative likelihood function as the contour function of a consonant belief function or, equivalently, as a possibility distribution [117, 30] is, to a large extent, consistent with statistical practice. For instance, likelihood intervals [42, 96] are focal intervals of the relative likelihood viewed as a possibility distribution. In the case where  $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$  is a vector parameter, the marginal contour function on  $\Theta_1$  is

$$pl_x(\theta_1) = \sup_{\theta_2 \in \Theta_2} pl_x(\theta_1, \theta_2), \quad (9.41)$$



which is the relative profile likelihood function when  $\theta_2$  is considered as a nuisance parameter. As another example, the usual likelihood ratio statistics  $\Lambda(x)$  for a composite hypothesis  $H_0 \subset \Theta$  can be seen as the plausibility of  $H_0$ , as

$$\Lambda(x) = \frac{\sup_{\theta \in H_0} L_x(\theta)}{\sup_{\theta' \in \Theta} L_x(\theta')} = \sup_{\theta \in H_0} pL_x(\theta) = Pl_x^\Theta(H_0). \quad (9.42)$$

However, the likelihood-based construction of belief functions is not compatible with Dempster's rule, i.e., Property (9.17) does not hold for this method. This does seem to be a weak point of this approach [80], although one might as well question Dempster's rule for combining independent statistical evidence, as was done by Aickin [1] and Walley [104], among others. Let  $E$  and  $E'$  be two independent random experiments with the same parameter space  $\Theta$ , producing outcomes  $x$  and  $y$  according to frequency distributions  $f_\theta(x)$  and  $g_\theta(y)$ . Let  $Bel_x^\Theta$  and  $Bel_y^\Theta$  denote the belief functions on  $\Theta$  obtained after observing  $x$  and  $y$ , respectively. It is clear that  $Bel_x^\Theta \oplus Bel_y^\Theta$  and  $Bel_{xy}^\Theta$  are different, although they have the same contour function. However,  $Bel_{xy}^\Theta$  can be obtained from  $Bel_x^\Theta$  and  $Bel_y^\Theta$  using the product rule of Possibility theory [30], which amounts to multiplying the contour functions (or possibility distributions) and renormalizing:

$$pl_{xy}(\theta) = \frac{pl_x(\theta)pl_y(\theta)}{\sup_{\theta' \in \Theta} pl_x(\theta')pl_y(\theta')}. \quad (9.43)$$

The apparent inadequacy of Dempster's rule in this case remains to be convincingly explained. It might be that different kinds of evidence require different combination mechanisms, as suggested by Dubois et al. in [26].

## 9.4 Prediction

In Section 9.2 and 9.3, we have described two solutions to the estimation problem, which consists in making statements about some parameter  $\theta$  after observing a realization  $x$  of some random quantity  $X \sim f_\theta(x)$ . The *prediction* problem considered in this section is, in some sense, the inverse of the previous one: given some knowledge about  $\theta$  obtained by observing  $x$  (represented here by a belief function), we wish to make statements about some random quantity  $Y \in \mathcal{Y}$  whose conditional distribution  $g_{x,\theta}(y)$  given  $X = x$  depends on  $\theta$ . In some cases,  $X = (X_1, \dots, X_n)$  and  $Y = X_{n+1}$ , where  $X_1, \dots, X_n, X_{n+1}$  is an iid sample. However, the model used here is more general. For instance,  $X = (Z_0, Z_1, \dots, Z_T)$  might be a time series and  $Y = (Z_{T+1}, \dots, Z_{T+h})$  might represent  $h$  future values to be predicted. For

simplicity, we will assume  $Y$  to be a one-dimensional random variable in the rest of this section. The method presented here was introduced in [51, 50].

### 9.4.1 General method

To make statements about  $Y$ , given some partial knowledge about  $\theta$ , we need to describe the relation between these two quantities. In the Dempster-Shafer framework, the uncertain relation between two variables is expressed by a joint belief function. Such a relation can be obtained by considering a  $\varphi$ -equation such as (9.9),

$$Y = \varphi'(\theta, W'), \quad (9.44)$$

where, as before, we will assume, without loss of generality,  $W'$  to have a standard uniform distribution. This equation defines a multi-valued mapping

$$\Gamma' : W \rightarrow \Gamma'(W) = \{(y, \theta) \in \mathcal{Y} \times \Theta \mid y = \varphi'(\theta, W')\}, \quad (9.45)$$

where  $\mathcal{Y}$  is the sample space of  $Y$ . The source  $([0, 1], \mathcal{B}([0, 1]), Pr_{W'}, \Gamma)$  defines a joint belief function  $Bel^{\mathcal{Y} \times \Theta}$  on  $\mathcal{Y} \times \Theta$ .

We now have two belief functions,  $Bel_x^\Theta$  and  $Bel^{\mathcal{Y} \times \Theta}$ , induced by multi-valued mapping  $S \rightarrow \Gamma_x(S)$  and  $W' \rightarrow \Gamma'(W')$ . Assuming the random variable  $S$  and  $W'$  to be independent, a belief function  $Bel^\mathcal{Y}$  on  $\mathcal{Y}$  can be obtained by combining  $Bel_x^\Theta$  and  $Bel^{\mathcal{Y} \times \Theta}$  using Dempster's rule and marginalizing on  $\mathcal{Y}$ ,

$$Bel_x^\mathcal{Y} = (Bel_x^\Theta \oplus Bel^{\mathcal{Y} \times \Theta}) \downarrow^\mathcal{Y}. \quad (9.46)$$

The corresponding multi-valued mapping is

$$\Gamma_\cap(S, W') = [(\mathcal{Y} \times \Gamma_x(S)) \cap \Gamma'(W')] \downarrow^\mathcal{Y} \quad (9.47a)$$

$$= \varphi'(\Gamma_x(S), W'). \quad (9.47b)$$

We then have, for any measurable  $A \subseteq \mathcal{Y}$ ,

$$Bel_x^\mathcal{Y}(A) = \mathbb{P}_{S, W'}(\varphi'(\Gamma_x(S), W') \subseteq A), \quad (9.48a)$$

$$Pl_x^\mathcal{Y}(A) = \mathbb{P}_{S, W'}(\varphi'(\Gamma_x(S), W') \cap A \neq \emptyset), \quad (9.48b)$$

where  $\mathbb{P}_{S, W'}$  is the product measure  $\mathbb{P}_S \otimes \mathbb{P}_{W'}$ .

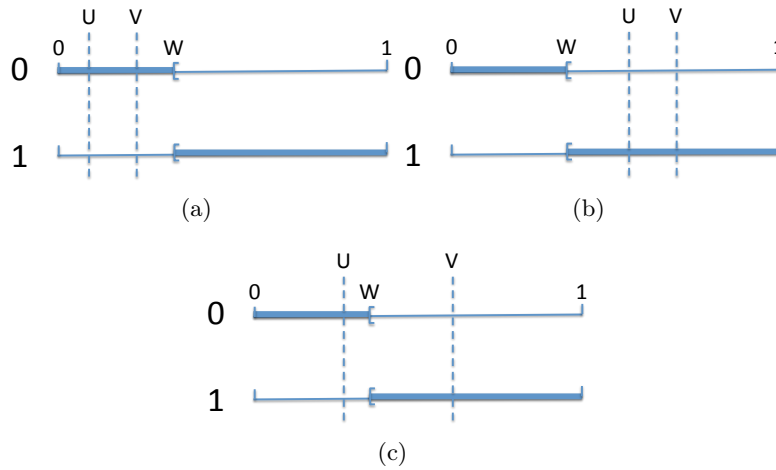


Figure 9.4: Three cases in the computation of the predictive belief function on  $Y$  in the Bernoulli example.

### 9.4.2 Bernoulli example

Let us return to the Bernoulli example considered in Sections 9.2.2 and 9.3.2 and let us address the problem of predicting a new value  $Y$  drawn independently from the Bernoulli  $\mathcal{B}(\theta)$  distribution. We have

$$Y = \varphi'(\theta, W') = \begin{cases} 1 & \text{if } W' \leq \theta \\ 0 & \text{otherwise,} \end{cases} \quad (9.49)$$

where  $W'$  has a standard uniform distribution.

Each of the two estimation methods studied in Sections 9.2 and 9.3 provides a random interval  $\Gamma_x(S) = [U(S), V(S)]$  on  $\Theta$ . Therefore, (9.47) becomes

$$[(\{0, 1\} \times [U, V]) \cap \Gamma(W')] \downarrow^{\{0,1\}} = \begin{cases} \{0\} & \text{if } [U, V] \subseteq [0, W'], \\ \{1\} & \text{if } [U, V] \subseteq [W', 1], \\ \{0, 1\} & \text{otherwise.} \end{cases} \quad (9.50)$$

The three cases in (9.50) are illustrated in Figures (9.4(a)), (9.4(b)) and

(9.4(c)). It follows from (9.50) that the predictive mass function  $m^{\mathcal{Y}}$  is

$$m^{\mathcal{Y}}(\{0\}) = P([U, V] \subseteq [0, W']) \quad (9.51a)$$

$$= \int P(V < W' | V = v) f(v) dv \quad (9.51b)$$

$$= \int (1 - v) f(v) dv = 1 - \mathbb{E}(V), \quad (9.51c)$$

$$m^{\mathcal{Y}}(\{1\}) = P([U, V] \subseteq [W', 1]) \quad (9.52a)$$

$$= \int P(U \geq W' | U = u) f(u) du \quad (9.52b)$$

$$= \int u f(u) du = \mathbb{E}(U), \quad (9.52c)$$

and

$$m^{\mathcal{Y}}(\{0, 1\}) = \mathbb{E}(V) - \mathbb{E}(U). \quad (9.53)$$

Equivalently, we have

$$Bel^{\mathcal{Y}}(\{1\}) = m^{\mathcal{Y}}(\{1\}) = \mathbb{E}(U) \quad (9.54)$$

and

$$Pl^{\mathcal{Y}}(\{1\}) = m^{\mathcal{Y}}(\{1\}) + m^{\mathcal{Y}}(\{0, 1\}) = \mathbb{E}(V). \quad (9.55)$$

When  $[U, V]$  is computed using Dempster's method, the expectations of  $U$  and  $V$  are given by (9.31). We then have

$$Bel^{\mathcal{Y}}(\{1\}) = \frac{N}{n+1}, \quad (9.56)$$

$$Pl^{\mathcal{Y}}(\{1\}) = \frac{N+1}{n+1}, \quad (9.57)$$

and the mass on  $\mathcal{Y} = \{0, 1\}$  is  $m(\{0, 1\}) = 1/(n+1)$ .

When  $[U, V]$  are computed using the likelihood-based approach, the predictive mass function has a more complicated expression. However, it has a very simple graphical representation in relation to the normalized likelihood function. As  $\mathbb{P}(U \geq 0) = \mathbb{P}(V \geq 0) = 1$ , we can write, using (7.57),

$$\mathbb{E}(U) = \int_0^{+\infty} (1 - F_U(u)) du \quad (9.58a)$$

$$= \int_0^{\hat{\theta}} (1 - pl_x(u)) du \quad (9.58b)$$

$$= \hat{\theta} - \int_0^{\hat{\theta}} pl_x(u) du \quad (9.58c)$$

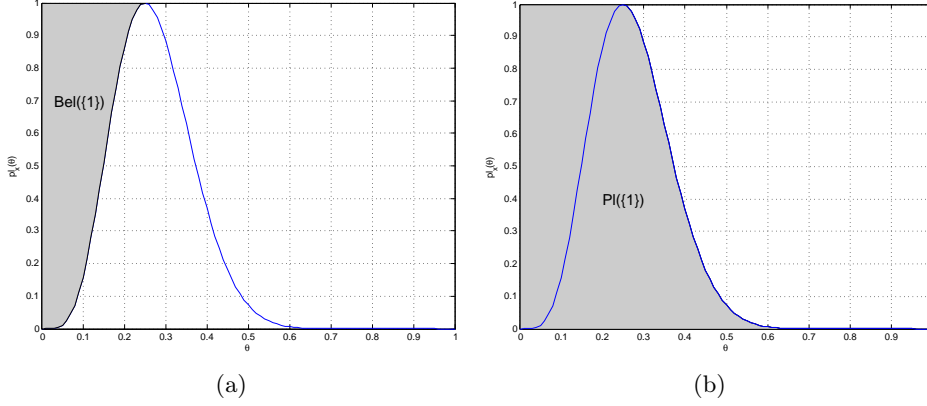


Figure 9.5: Predictive belief and plausibility of success for a Bernoulli trial based on the contour function  $pl_x(\theta)$  on the probability of success  $\theta$ .

and

$$\mathbb{E}(V) = \int_0^{+\infty} (1 - F_V(v)) dv \quad (9.58d)$$

$$= \hat{\theta} + \int_{\hat{\theta}}^1 pl_x(v) dv. \quad (9.58e)$$

These two quantities can be represented as the areas of regions delimited by the contour function, as shown in Figure 9.5. The difference  $Pl_x^{\mathcal{Y}}(\{1\}) - Bel_x^{\mathcal{Y}}(\{1\})$ , which is the mass  $m_x^{\mathcal{Y}}(\{0, 1\})$  assigned to ignorance, is simply the area under the contour function  $pl_x$ . It tends to zero as the sample size  $n$  tends to infinity.

### 9.4.3 Monte Carlo approximation

In practice, the analytical expression of the predictive belief function induced by the multi-valued mapping (9.47) may be intractable. We can then resort to Monte Carlo approximation, by drawing independently  $n$  pairs  $(S_1, W_1), \dots, (S_n, W_n')$  from the continuous uniform distribution in  $[0, 1]^2$ . For any measurable  $A \subseteq \mathbb{R}$ , we may then approximate  $Bel_x^{\mathcal{Y}}(A)$  and  $Pl_x^{\mathcal{Y}}(A)$  in (9.48) by, respectively,

$$\widehat{Bel}_x^{\mathcal{Y}}(A) = \frac{1}{n} \#\{\varphi'(\Gamma_x(S_i), W_i') \subseteq A\}, \quad (9.59a)$$

$$\widehat{Pl}_x^{\mathcal{Y}}(A) = \frac{1}{n} \#\{\varphi'(\Gamma_x(S_i), W_i') \cap A \neq \emptyset\}. \quad (9.59b)$$

If the mapping  $\theta \rightarrow \varphi'(\theta, w)$  is continuous for any  $w$  and if the sets  $\Gamma_x(S_i)$  are closed and convex, then  $\varphi'(\Gamma_x(S_i), W'_i)$  is an interval  $[a_i, b_i]$ , whose bounds can be computed by solving the following optimization problems:

$$a_i = \min \varphi'(\theta, W_i) \quad (9.60a)$$

and

$$b_i = \max \varphi'(\theta, W_i) \quad (9.60b)$$

under the constraint

$$L_x(\theta) \geq S_i. \quad (9.60c)$$

The lower and upper expectations of  $Y$  with respect to the predictive belief function can then be approximated by the mean of the  $a_i$ 's and  $b_i$ 's,

$$\mathbb{E}_*(Y) \approx \frac{1}{n} \sum_{i=1}^n a_i, \quad \mathbb{E}^*(Y) \approx \frac{1}{n} \sum_{i=1}^n b_i. \quad (9.61)$$

Similarly, the lower and lower predictive quantiles of  $Y$  (see Section 8.2.2) can be approximated by the empirical quantiles of the  $a_i$ 's and  $b_i$ 's.

#### 9.4.4 Relationship with the Bayesian posterior predictive distribution

To conclude this section, we can remark that the predictive belief function  $Bel_x^{\mathcal{Y}}$  boils down to the Bayesian posterior predictive distribution of  $Y$  given  $X = x$  when a prior probability distribution  $\pi(\theta)$  is available and combined with the belief function  $Bel_x^{\Theta}$  by Dempster's rule. As mentioned in Section 9.3.1, the combined belief function  $Bel_x^{\Theta} \oplus \pi$  is then, by construction, the posterior probability distribution  $f_x(\theta)$  of  $\theta$  given  $X = x$  and we then have, for any measurable subset  $A \subseteq \mathcal{Y}$ :

$$Bel_x^{\mathcal{Y}}(A) = \mathbb{P}(\varphi'(\theta, W') \in A|x) \quad (9.62a)$$

$$= \int_{\Theta} \mathbb{P}(\varphi'(\theta, W') \in A|\theta, x) f_x(\theta) d\theta \quad (9.62b)$$

$$= \int_{\Theta} \left( \int_A g_{x,\theta}(y) dy \right) f_x(\theta) d\theta \quad (9.62c)$$

$$= \int_A \left( \int_{\Theta} g_{x,\theta}(y) f_x(\theta) d\theta \right) dy \quad (9.62d)$$

$$= \int_A g_x(y) dy, \quad (9.62e)$$

which is the posterior predictive probability that  $Y$  belongs to  $A$ , given  $x$ .

The forecasting method introduced in this chapter is thus a proper generalization of the Bayesian approach. The two methods coincide when a prior probability distribution of the parameter is provided. However, this is not required in the belief function approach, making it less arbitrary than the Bayesian approach in the absence of prior knowledge about the data distribution.





## Chapter 10

# Uncertain Data



## Chapter 11

# Classification



## Chapter 12

# Clustering



## Chapter 13

# Belief functions on sets of relations





# Bibliography

- [1] M. Aickin. Connecting Dempster-Shafer belief functions with likelihood-based inference. *Synthese*, 123:347–364, 2000.
- [2] J. Aldrich. R. A. Fisher and the making of the maximum likelihood 1912-1922. *Statistical Science*, 12:162–176, 1997.
- [3] M. Allais. Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école Américaine. *Econometrica*, 21(4):503–546, 1953.
- [4] R. G. Almond. *Graphical belief models*. Chapman and Hall, London, 1995.
- [5] D. A. Alvarez. On the calculation of the bounds of probability of events using infinite random sets. *International Journal of Approximate Reasoning*, 43(3):241–267, 2006.
- [6] K. J. Arrow and L. Hurwicz. An optimality criterion for decision making under ignorance. In K. J. Arrow and L. Hurwicz, editors, *Studies in Resource Allocation Processes*. Cambridge University Press, 1977.
- [7] G. A. Barnard, G. M. Jenkins, and C. B. Winsten. Likelihood inference and time series. *Journal of the Royal Statistical Society*, 125(3):321–372, 1962.
- [8] J. A. Barnett. Computational methods for a mathematical theory of evidence. In *Proceedings IJCAI-81*, pages 868–875, Vancouver, 1981.
- [9] M. Bauer. Approximation algorithms and decision making in the Dempster-Shafer theory of evidence – an empirical study. *International Journal of Approximate Reasoning*, 17:217–237, 1997.

- [10] J. O. Berger and R. L. Wolpert. *The likelihood principle: a review, generalizations, and statistical implications*, volume 6 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 2nd edition, 1988.
- [11] J. M. Bernardo and A. F. M. Smith. *Bayesian theory*. John Wiley and Sons, New-York, 1994.
- [12] A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.
- [13] G. Choquet. Théorie des capacités. *Annales de l’Institut Fourier (Grenoble)*, V:131–295, 1953.
- [14] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–12, 1946.
- [15] B. de Finetti. La prévision : ses lois logiques, ses sources objectives. *Annales de l’Institut Henri Poincaré*, 7(1):1–68, 1937.
- [16] M. A. B. Deakin. The wine/water paradox: background, provenance and proposed resolutions. *The Australian Mathematical Society Gazette*, 33(3):200–205, 2006.
- [17] A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374, 1966.
- [18] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.
- [19] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 57:515–528, 1967.
- [20] A. P. Dempster. A generalization of Bayesian inference (with discussion). *J. R. Statistical Society B*, 30:205–247, 1968.
- [21] A. P. Dempster. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics*, 39(3):957–966, 1968.
- [22] A. P. Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.

- [23] T. Denœux. Inner and outer approximation of belief structures using a hierarchical clustering approach. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(4):437–460, 2001.
- [24] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.
- [25] T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.
- [26] D. Dubois, S. Moral, and H. Prade. Belief change rules in ordinal and numerical uncertainty theories. In D. Dubois and H. Prade, editors, *Belief change*, volume 3 of *Handbook of defeasible reasoning and uncertainty management systems*, pages 311–392. Kluwer Academic Publishers, Boston, 1998.
- [27] D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In M. M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-Holland, New-York, 1982.
- [28] D. Dubois and H. Prade. A note on measures of specificity for fuzzy sets. *International Journal of General Systems*, 10(4):279–283, 1985.
- [29] D. Dubois and H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12(3):193–226, 1986.
- [30] D. Dubois and H. Prade. *Possibility Theory: An approach to computerized processing of uncertainty*. Plenum Press, New-York, 1988.
- [31] A. W. F. Edwards. *Likelihood (expanded edition)*. The John Hopkins University Press, Baltimore, USA, 1992.
- [32] D. Ellsberg. Risk, ambiguity and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.
- [33] R. A. Fisher. On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.
- [34] R. A. Fisher. *Statistical methods and scientific inference*. Oliver and Boyd, Edinburgh, 1956.

- [35] P. Ghirardato. Coping with ignorance: unforeseen contingencies and non-additive uncertainty. *Economic theory*, 17:247–276, 2001.
- [36] P. H. Giang. Decision with Dempster-Shafer belief functions: Decision under ignorance and sequential consistency. *International Journal of Approximate Reasoning*, 53(1):38 – 53, 2012.
- [37] P. H. Giang and P. P. Shenoy. A decision theory for partially consonant belief functions. *International Journal of Approximate Reasoning*, 52(3):375–394, 2011.
- [38] I. Gilboa and D. Schmeidler. Additive representations of non-additive measures and the choquet integral. *Annals of Operations Research*, 51:43–65, 1994.
- [39] F. Gul. Savage’s theorem with a finite number of states. *Journal of Economic Theory*, 57(1):99–110, 1992.
- [40] D. Harmanec. Faithful approximations of belief functions. In K. B. Laskey and H. Prade, editors, *Uncertainty in Artificial Intelligence 15 (UAI99)*, Stockholm, 1999.
- [41] D. Harmanec and G. J. Klir. Measuring total uncertainty in Dempster-Shafer theory: A novel approach. *International Journal of General Systems*, 22(4):405–419, 1994.
- [42] D. J. Hudson. Interval estimation from the likelihood function. *J. R. Statistical Society B*, 33(2):256–262, 1973.
- [43] J.-Y. Jaffray. Application of linear utility theory to belief functions. In B. Bouchon, L. Saitta, and R. Yager, editors, *Uncertainty and Intelligent Systems*, volume 313 of *Lecture Notes in Computer Science*, pages 1–8. Springer Berlin Heidelberg, 1988.
- [44] J.-Y. Jaffray. Linear utility theory for belief functions. *Operations Research Letters*, 8(2):107 – 112, 1989.
- [45] J.-Y. Jaffray and P. Wakker. Decision making with belief functions: compatibility and incompatibility with the sure-thing principle. *Journal of Risk and Uncertainty*, 8:255–271, 1994.
- [46] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*. Springer-Verlag, London, 2001.

- [47] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003.
- [48] R. Jeffrey. *The Logic of Decision*. McGraw-Hill, New York, 1965.
- [49] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.
- [50] O. Kanjanatarakul, P. Lertpongpiroon, S. Singkharat, and S. Sriboonchitta. Econometric forecasting using linear regression and belief functions. In F. Cuzzolin, editor, *Belief functions: theory and applications 3*, Advances in Intelligent and Soft Computing, Oxford, UK, 2014. Springer.
- [51] O. Kanjanatarakul, S. Sriboonchitta, and T. Denceux. Forecasting using belief functions: an application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.
- [52] R. Kennes. Computational aspects of the Möbius transform of graphs. *IEEE Trans. SMC*, 22:201–223, 1992.
- [53] F. Klawonn and P. Smets. The dynamic of belief in the transferable belief model and specialization-generalization matrices. In D. Dubois, M. Wellman, B. D’Ambrosio, and P. Smets, editors, *Proc. of the 8th conference on Uncertainty in Artificial Intelligence*, pages 130–137. Morgan Kaufmann, San Mateo, CA, 1992.
- [54] G. J. Klir and M. J. Wierman. *Uncertainty-Based Information. Elements of Generalized Information Theory*. Springer-Verlag, New-York, 1999.
- [55] J. Kohlas and P.-A. Monney. *A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory of Evidence*. Springer-Verlag, Berlin, 1995.
- [56] J. Kohlas and P.-A. Monney. An algebraic theory for statistical information based on the theory of hints. *International Journal of Approximate Reasoning*, 48(2):Pages 378–398, 2008.
- [57] D. V. Lindley. *Understanding Uncertainty*. John Wiley & Sons, 2006.
- [58] J. D. Lowrance, T. D. Garvey, and T. M. Strat. A framework for evidential-reasoning systems. In T. K. et al., editor, *Proceedings of AAAI’86*, volume 2, pages 896–903, Philadelphia, August 1986. AAAI.

- [59] R. D. Luce and H. Raiffa. *Games and Decisions: Introduction and Critical Survey*. Wiley, New York, 1957.
- [60] Y. Maeda and H. Ichihashi. An uncertainty measure with monotonicity under the random set inclusion. *International Journal of General Systems*, 21(4):379–392, 1993.
- [61] J. M. Mikkelsen. Dissolving the wine/water paradox. *British Journal for the Philosophy of Science*, 55(1):137–145, 2004.
- [62] P.-A. Monney. *A Mathematical Theory of Arguments for Statistical Evidence*. Contributions to Statistics. Physica-Verlag, Heidelberg, 2003.
- [63] R. E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New-Jersey, 1966.
- [64] R. E. Moore, R. B. Kearfott, and M. J. Cloud. *Introduction to Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, 2009.
- [65] S. Moral and N. Wilson. Markov-chain Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In *Proc. of the Twelfth National Conference on Artificial intelligence (AAAI-94)*, volume 1, pages 269–274, 1994.
- [66] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.
- [67] H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC Press, Boca Raton, Florida, 2006.
- [68] Z. Pawlak. Rough sets. *International Journal of Parallel Programming*, 11(5):341–356, 1982.
- [69] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishing, Dordrecht, 1991.
- [70] B. W. Pearson, III. *Introduction to Management Science*. Prentice Hall, Upper Saddle River, NJ, USA, 12th edition, 2014.
- [71] K. R. Popper. The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:25–42, 1959.
- [72] A. Ramer. Uniqueness of information measure in the theory of evidence. *Fuzzy Sets and Systems*, 24:183–196, 1987.

- [73] F. P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter VII, pages 156–198. Harcourt, Brace and Company, New York, 1931.
- [74] L. J. Savage. *The foundations of statistics*. Wiley, New York, 1954.
- [75] D. Schmeidler. Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97:255–261, 1986.
- [76] G. Shafer. *Allocations of probability: A theory of partial belief*. PhD thesis, Princeton University, 1973.
- [77] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
- [78] G. Shafer. Allocations of probability. *Annals of Probability*, 7(5):827–839, 1979.
- [79] G. Shafer. Constructive probability. *Synthese*, 48(1):1–60, 1981.
- [80] G. Shafer. Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B*, 44:322–352, 1982.
- [81] G. Shafer. Savage revisited (with discussion). *Statistical Science*, 1:463–501, 1986.
- [82] G. Shafer. Comments on “Constructing a logic of plausible inference: a guide to Cox’s Theorem”, by Kevin S. Van Horn. *International Journal of Approximate Reasoning*, 35(1):97–105, 2004.
- [83] G. Shafer, P. P. Shenoy, and K. Mellouli. Propagating belief functions in qualitative Markov trees. *International Journal of Approximate Reasoning*, 1:349–400, 1987.
- [84] L. S. Shapley. A value for  $n$ -person games. In H. Kuhn and A. W. Tucker, editors, *Contributions to the theory of games*, volume 2, pages 307–317. Princeton University Press, Princeton, NJ, US, 1953.
- [85] P. P. Shenoy. A valuation-based language for expert systems. *International Journal of Approximate Reasoning*, 3:383–411, 1989.
- [86] P. P. Shenoy. Binary joint trees for computing marginals in the Shenoy-Shafer architecture. *International Journal of Approximate Reasoning*, 17:239–263, 1997.

- [87] P. Smets. Information content of an evidence. *International Journal of Man-Machine Studies*, 19:33–43, 1983.
- [88] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, R. D. Schachter, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 29–40. North-Holland, Amsterdam, 1990.
- [89] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.
- [90] P. Smets. The canonical decomposition of a weighted belief. In *Int. Joint Conf. on Artificial Intelligence*, pages 1896–1901, San Mateo, Ca, 1995. Morgan Kaufman.
- [91] P. Smets. The application of the matrix calculus to belief functions. *International Journal of Approximate Reasoning*, 31(1–2):1–30, 2002.
- [92] P. Smets. Decision making in a context where uncertainty is represented by belief functions. In R. P. Srivastava and T. J. Mock, editors, *Belief functions in business decisions*, pages 17–61. Physica-Verlag, Heidelberg, 2002.
- [93] P. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38:133–147, 2005.
- [94] P. Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8(4):387–412, 2007.
- [95] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
- [96] D. A. Sprott. *Statistical Inference in Science*. Springer-Verlag, Berlin, 2000.
- [97] R. P. Srivastava. Decision making under ambiguity: A belief-function perspective. *Archives of Control Sciences*, 6 (XLII)(1–2):5–27, 1997.
- [98] T. M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4(5–6):391–417, 1990.
- [99] K. Szaniawski. Some remarks concerning the criterion of rational decision making. *Studia Logica*, 9(1):221–239, 1960.



- [100] B. Tessem. Approximations for efficient computation in the theory of evidence. *Artificial Intelligence*, 61:315–329, 1993.
- [101] M. C. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17 – 29, 2007.
- [102] K. S. Van Horn. Constructing a logic of plausible inference: a guide to Cox’s theorem. *International Journal of Approximate Reasoning*, 34(1):3–24, 2003.
- [103] J. von Neumann and O. Morgenstern. *Theory Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.
- [104] P. Walley. Belief function representations of statistical evidence. *The Annals of Statistics*, 15(4):1439–1465, 1987.
- [105] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- [106] L. A. Wasserman. Belief functions and statistical evidence. *The Canadian Journal of Statistics*, 18(3):183–196, 1990.
- [107] N. Wilson. Algorithms for Dempster-Shafer theory. In D. M. Gabbay and P. Smets, editors, *Handbook of defeasible reasoning and uncertainty management. Volume 5: Algorithms for Uncertainty and Defeasible Reasoning*, pages 421–475. Kluwer Academic Publishers, Boston, 2000.
- [108] W. Xiong and H. Liu. An axiomatic foundation for Yager’s decision theory. *International Journal of Intelligent Systems*, 29(4):365–387, 2014.
- [109] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denceux. Multi-modal information for urban scene understanding. *Machine Vision and Applications*, 27:331–349, 2016.
- [110] R. R. Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of general Systems*, 9:249–260, 1983.
- [111] R. R. Yager. The entailment principle for Dempster-Shafer granules. *Int. J. of Intelligent Systems*, 1:247–262, 1986.
- [112] R. R. Yager. On ordered weighted averaging aggregation operators in multi- criteria decision making. *IEEE Trans. on Systems, Man and Cybernetics*, 18:83–190, 1988.

- [113] R. R. Yager. Decision making under Dempster-Shafer uncertainties. *International Journal of General Systems*, 20(3):233–245, 1992.
- [114] R. R. Yager. Decision making using minimization of regret. *International Journal of Approximate Reasoning*, 36(2):109–128, 2004.
- [115] Y. Y. Yao and P. J. Lingras. Interpretations of belief functions in the theory of rough sets. *Information Sciences*, 104(1–2):81–106, 1998.
- [116] L. A. Zadeh. Fuzzy sets. *Inform. Control*, 8:338–353, 1965.
- [117] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

# Index

- allocation, [42](#)
- axioms
  - Cox, [26](#)
  - Savage, [26](#), [97](#)
- contour function, [40](#)
- credal set, [42](#), [102](#)
- mass function, [31](#)
  - Bayesian, [32](#)
  - consonant, [40](#)
  - logical, [32](#)
  - separable, [57](#)
  - simple, [57](#)
- orthogonal sum, [48](#)
- paradox
  - Ellsberg's, [99](#), [103](#)
  - wine/water, [28](#)
- probability, [22](#)
  - compatible, [102](#)
  - imprecise, [42](#)
  - objective, [22](#)
  - subjective, [22](#)
- relation
  - preference, [90](#)
- Sure Thing Principle, [98](#)
- utility, [89](#)