Introduction to the theory of belief functions

Thierry Denœux

August 4, 2015

Contents

List of Figures

List of Tables

Chapter 1

Uncertainty

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 [15, 17]. It was then further developed by Glenn Shafer [58] 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 Demspter-Shafer theory, has been widely used in several areas such as Artificial Intelligence, Information Fusion and Risk Analysis. Recently, there has been a revived interested in its application to statistical inference. This formalism seems particularly well suited to situations where we are facing limited information such as uncertain and low quality data, partially reliable and conflicting expert opinions, or both. There has been thousands of applications in many domains, including engineering, medicine, economics, 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 Sources of uncertainty

Uncertainty is ubiquitous in every area of human activity. Typically, we are interested in some question Q, such as: What is the mean value of some variable in a population? What will be the economic growth rate in the

United States next year? What was the amount of carbon dioxide emission in China in 2012? etc. In the following, we will denote by Ω the set of possible answers (one and only one is assumed to be true), and by $\boldsymbol{\omega}$ the true answer. If we know the exact value of $\boldsymbol{\omega}$, this is a situation of complete certainty. If we know nothing at all (except that $\boldsymbol{\omega}$ is in Ω), we have complete uncertainty. Actually, these two extreme situations are not frequent: usually, we have only partial knowledge of $\boldsymbol{\omega}$, based on limited evidence about the question of interest. The issue then arises of how to represent such partial information in such a way that it can be used for further reasoning, computation and rational decision making.

It has become customary in some areas (such as risk analysis) to distinguish between two main sources of uncertainty:

1. When the question of interest concerns some property of an object taken at random from a well-defined population (such as, e.g., the color of a ball to be drawn from an urn), we say that we have *random*, *aleatory* or *physical* uncertainty. Such uncertainty cannot be reduced because it depends on the physical property of the population and of the random experiment.

2. In many situations, uncertainty does not arise from randomness but from lack of knowledge. For instance, the name of the next president of the US is unknown, but it is not random because there is no notion of random experiment (in particular, the next presidential election will occur only once in precisely the same context). Such uncertainty is said to be *epistemic*. It can be reduced by acquiring further information related to the question of interest.

The two main classical formalisms for representing uncertainty are the set theory (or, equivalently, propositional logic) and probability theory. These approaches will be discussed below, with greater emphasis on probability theory, which is by far the most widely used framework.

1.2 Set-based representation of uncertainty

Perhaps the simplest way of representing partial knowledge about some question is as a set $A \subseteq \Omega$ that certainly contains the true answer $\boldsymbol{\omega}$. There is a vast literature on set-membership approaches to uncertainty, with application, e.g., in computer science and automatic control [47, 48, 33]. An important special case is interval arithmetics, which includes 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 5. We will briefly review these operations in the following section, before showing the link with propositional logic. 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 Ω, assumed to contain the answer to the question of interest. How to combine these pieces of information?

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 Ω 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 rules 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 Ω that belong to A or 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.

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 Ω_X and Ω_Y be the sets of possible values 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 Ω_{XY}, and defined as the set of ordered pairs (x, y) of an element of Ω_X and an element of Ω_y. A subset of R of Ω_{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.*

Let R be a relation on Ω_{XY}. The *projection* of R onto Ω_X, denoted by

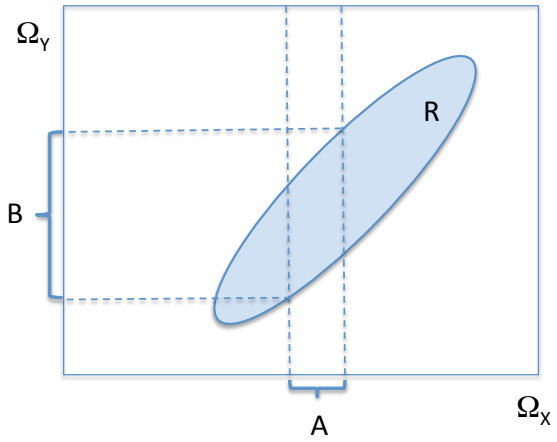


Figure 1.1: Reasoning with sets. Knowing that $X \in A$ and the relation R constraining the values of X and Y, we can deduce that $Y \in B$, where B is the projection on Ω_X of the intersection of R with the cylindrical extension of A.

$R \downarrow \Omega_X$, is the subset of Ω_X defined by

$$R \downarrow \Omega_X = \{x \in \Omega_X | \exists y \in \Omega_Y, (x, y) \in R\}. \tag{1.1}$$

Symmetrically,

$$R \downarrow \Omega_Y = \{y \in \Omega_Y | \exists x \in \Omega_X, (x, y) \in R\}. \tag{1.2}$$

Conversely, let A be a subset of Ω_X. Its *cylindrical extension* in Ω_{XY}, denoted by $A \uparrow \Omega_{XY}$, is the subset of Ω_{XY} defined as

$$A \uparrow \Omega_{XY} = A \times \Omega_Y = \{(x, y) \in \Omega_{XY} | x \in A\}. \tag{1.3}$$

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 Ω_X;
- Evidence about the values that X and Y can take jointly, represented by a relation $R \subseteq \Omega_{XY}$ (see Figure 1.1).

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 of there is some x in A such

that $(x, y) \in R$. Formally:

$$B = \{y \in \Omega_Y | \exists x \in A, (x, y) \in R\}, \tag{1.4}$$

which can be written as

$$B = (R \cap (A \uparrow \Omega_{XY})) \downarrow \Omega_Y. \tag{1.5}$$

This kind of reasoning may straightforwardly be applied to any number of variables. As we will see in Chapter 5, 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* $p, q, r, \ldots$, which represent statements 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

statest 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 ϕ 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 ϕ and ψ are two formulas, then $\mathcal{I}(\phi \wedge \psi) = \mathcal{I}(\phi) \cap \mathcal{I}(\phi)$, $\mathcal{I}(\phi \vee \psi) = \mathcal{I}(\phi) \cup \mathcal{I}(\phi)$ 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. Propositional logic and set theory thus have the same expressive power.

1.2.3 Limitations of sets for representing uncertainty

The main limitation of set-based representations of uncertainty (and propositional logic) is that 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 contain the true value with full certainty. A lot of information is usually lost in such a representation. For instance, if an expert is asked to give an interval that surely contains the mean sea level in 2050, he will give a wide interval, even though he may actually believe that the mean sea level will be contained within narrower bounds. As we will show later, belief functions can be seen as extending the notion of set by allowing one to provide different sets with attached degrees of support.

1.3 Probabilistic representation of uncertainty

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

Let Ω be a set and $\mathcal{A} \subseteq 2^\Omega$ an *algebra* of subsets of Ω, defined as non-empty collection of subsets of Ω (called *events*), closed under complementation and finite union, i.e., for all A and B in $\mathcal{A}$, $A \cup B \in \mathcal{A}$. We can remark that Ω necessarily belongs to $\mathcal{A}$. A *finitely additive probability measure* on $(\Omega, \mathcal{A})$ is a function P from $\mathcal{A}$ to $[0, 1]$ such that

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

$$P(A \cup B) = P(A) + P(B). \tag{1.6}$$

We can easily deduce from (1.6) that, for any elements A and B of $\mathcal{A}$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.7}$$

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

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} P\left(\bigcap_{i \in I} A_i\right). \tag{1.8}$$

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 Ω that is closed under complementation and countable union. Such a family is called a *σ-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), $i = 1, \ldots, \infty$, of pairwise disjoint elements of $\mathcal{A}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{1.9}$$

The triple $(\Omega, \mathcal{A}, P)$ is called a *probability space*.

It is clear that the notions of finitely additive and countably additive probability measures differ only when the space Ω is infinite.

1.3.2 Interpretations

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 $n \to +\infty$), or as a *propensity* [53] (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 (1.9) 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., [9, 34]) 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 Ω to a random experiment with known chances [60]. This allows us a 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 $1/2$.

Behavioral approach In the behavioral approach, we assume that the belief state of an agent can be deduced from observing its betting behavior. The following "Dutch book" argument[1], first put forward by Ramsey [55] and de Finetti [12], shows that consistent betting behavior, in some sense, should be based on probabilities. 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.

[1] A Dutch book is a set of odds and bets which guarantees a profit, regardless of the outcome of the gamble.

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 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

If we interpret degrees of belief as betting rates, it can thus be argued that degrees of belief should be (finitely) additive and our state of knowledge should be represented by a probability measure. However, this point of view is open to criticism:

1. First, the betting scheme just described is a highly idealized situation, and it is debatable if any situation of choice under uncertainty can fit this idealized picture (probabilities and utilities do not exist, they are a construction). Additionally, it is not obvious that the setting of betting rates in some particular betting scheme is the primary purpose of probability judgement [60].

2. Secondly, by slightly changing the story, we can arrive at different conclusions. For instance, assume that you are not obliged to enter the game and you are not required to accept to be the banker. Let $P_*(A)$ be the highest price you are willing to pay for the lottery ticket. Then, a Dutch book can be raised against you iff P_* fails to be a lower

probability function, i.e., the lower envelope of a family of probability measures [77].

1.3.3 Cox axioms

Some scholars have attempted to justify the use of probabilities to represent degrees of belief using an axiomatic approach. In particular, the axioms of Cox [11] and Savage [56] 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 axioms. Savage's axioms 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 Ω. Consider the following axioms:

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

$$Cr(\overline{A}|B) = S[Cr(A|B)], \tag{1.10}$$

where S is a twice differentiable function;

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

$$Cr(A \cap A'|B) = F[Cr(A'|A \cap B), Cr(A|B)], \tag{1.11}$$

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

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

$$g[Cr(A|B)] \cdot g[Cr(B)] = g[Cr(A \cap B)] \tag{1.12}$$

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, $\overline{A}$ may be supported or not), and $Cr(A|\Omega)$ will not be determined by $Cr(\overline{A}|\Omega)$.

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 come back, given that he has already reached the point, and the probability that he can reach the point. Yet, as noted by Shafer, 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.

1.3.4 Two paradoxes

As shown in the previous section, attempts to justify the use of probabilities to represent degrees of belief have not settled the question. In contrast, there appears to be some serious arguments against the use of probability theory as a model of epistemic uncertainty (Bayesian model) In particular, the use of a probability distribution to represent ignorance may lead to some inconsistencies, and probability theory does not seem to be a plausible model of how people make decisions based on weak information. These arguments are exemplified by the following two paradoxes.

The wine/water paradox Assume that all we now about some quantity X is that it belongs to some set A. According to Laplace's principle of indifference (PI) – and also according to the principle of maximal entropy, this state of knowledge should be represented by assigning equal probabilities to any possible values of X. However, consider the following paradox, attributed to Von Mises (see [45] and [13] for recent reviews and discussions).

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:

$$P(X \leq 2) = (2 - 1/3)/(3 - 1/3) = 5/8. \tag{1.13}$$

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:

$$P(Y \geq 1/2) = (3 - 1/2)/(3 - 1/3) = 15/16. \tag{1.14}$$

By comparing (1.13) and (1.14), 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.

Ellsberg's paradox The following paradox is due to Ellsberg [28]. 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, hence $P(\text{red}) > P(\text{black})$, but they strictly prefer f_4 to f_3, hence $P(\text{black}) > P(\text{red})$.

This famous paradox shows that probability theory is not a plausible descriptive model of how people make decisions under ambiguity (i.e., when objective probabilities are not given).

1.4 Conclusions

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

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 6.

2.1 Mass function

2.1.1 Definitions

Let Ω 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 $\boldsymbol{\omega}$, and an arbitrary element of Ω by ω. Shafer [58] 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, Ω 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 5 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^Ω to the interval $[0, 1]$, such that $m(\emptyset) = 0$ and

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

As will be discussed later, each number $m(A)$ represents the probability that the evidence supports exactly the proposition $\boldsymbol{\omega} \in A$, 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
$m_?(A)$	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 Ω 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 Ω: 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 $m_?$ defined by $m_?(\Omega) = 1$; 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 $p : \Omega \to [0, 1]$ such that $p(\omega) = m(\{\omega\})$ for all $\omega \in \Omega$.

Example 2.1 *Consider the mass on functions on $\Omega = \{a, b, c\}$ shown in Table 2.1. Mass function m_1 is Bayesian, m_2 is logical, $m_?$ 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 5, 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 $\Omega = \{Peter, John, Mary\}$. 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 $\{Peter, John\}$, and 0.2 to Ω:

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

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 $\boldsymbol{\omega} \in \{\text{Peter}, \text{John}\}$, otherwise the meaning is $\boldsymbol{\omega} \in \Omega$, i.e., the message is totally uninformative.

As remarked by Shafer [60], 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.

More precisely, 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 $\boldsymbol{\omega} \in A$ 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

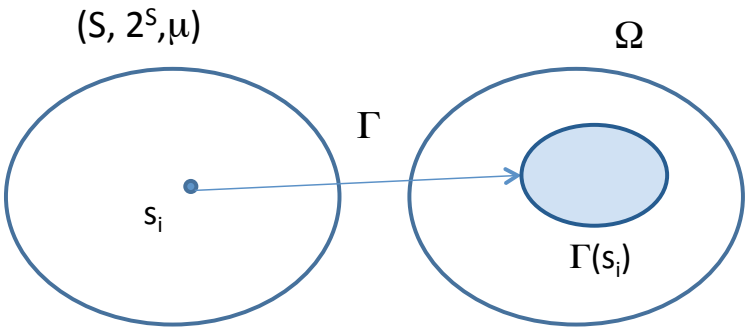


Figure 2.1: Random code setup.

function $m(A) = p$ and $m(\Omega) = 1 - p$. Such a mass function, with two focal sets including Ω, 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 $\boldsymbol{\omega} \in A$ 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 [60], 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 $\boldsymbol{\omega} \in A^*$ 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 $S = \{s_1, \ldots, s_r\}$, according to some known probability measure μ (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 $\boldsymbol{\omega} \in \Gamma(s)$ for some subset $\Gamma(s)$ of Ω. Then,

$$m(A) = \mu(\{s \in S | \Gamma(s) = A\}) \tag{2.2}$$

is the chance that the original message was "$\boldsymbol{\omega} \in A$", i.e., the *probability of knowing that* $\boldsymbol{\omega} \in A$, and nothing more.

In the above framework, the mapping $\Gamma : S \to 2^\Omega \setminus \{\emptyset\}$ 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 [51]. 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 Ω 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 (2.1). Conversely, any mass function can be seen as generated by a source. For instance, if $A_1, \ldots, A_n$ are the focal sets of a mass function m, we may set $S = \{1, \ldots, n\}$ and $\mu(\{i\}) = m(A_i)$ for $1 \leq i \leq n$. However, as we shall see in Chapter 6, 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 Ω generated by a source $(S, 2^S, \mu, \Gamma)$. For any $A \subseteq \Omega$, the uncertainty pertaining to the proposition $\boldsymbol{\omega} \in A$ can be quantified by two numbers:

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

$$\begin{aligned} Bel(A) &= \mu(\{s \in S | \Gamma(s) \subseteq A\}) && \text{(2.3a)} \\ &= \sum_{B \subseteq A} m(B); && \text{(2.3b)} \end{aligned}$$

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

$$\begin{aligned} Pl(A) &= \mu(\{s \in S | \Gamma(s) \cap A \neq \emptyset\}) && \text{(2.4a)} \\ &= \sum_{B \cap A \neq \emptyset} m(B). && \text{(2.4b)} \end{aligned}$$

Clearly, $Bel(\emptyset) = Pl(\emptyset) = 0$, $Bel(\Omega) = Pl(\Omega) = 1$, $Bel(A) \leq Pl(A)$ and $Pl(A) = 1 - Bel(\overline{A})$, where $\overline{A}$ 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 $Bel : 2^\Omega \to [0, 1]$ 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 $Pl : 2^\Omega \to [0, 1]$ 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 focals sets B_i, $i = 1, 2, 3, 4$. The degree of belief*

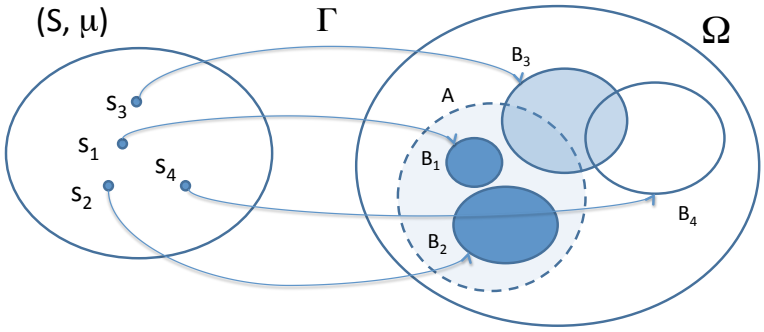


Figure 2.2: Belief and plausibility functions (Example 2.3).

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 $Bel(\overline{A}) = m(B_4)$, which is clearly equal to $1 - Pl(A)$.

2.2.2 Properties

Theorem 2.1 *A function $Bel : 2^\Omega \to [0, 1]$ 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 $A_1, \ldots, A_k$ of subsets of Ω,*

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

Proof: See [58, page 51].

Property (2.5) is a weaker form of the corresponding property (1.8), 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 Ω 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

$$Bel(B) = Bel(A \cup C) \geq Bel(A) + Bel(C) - Bel(A \cap C) \\ = Bel(A) + Bel(C) \geq Bel(A). \quad (2.6)$$

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 [58].

Theorem 2.2 *Let $Bel : 2^\Omega \to [0, 1]$ 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 [58, page 52].

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

Theorem 2.3 *A function $Pl : 2^\Omega \to [0, 1]$ 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 $A_1, \ldots, A_k$ of subsets of Ω,*

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

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

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

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 Ω have been arranged in some predefined order [68]. One such order with nice properties is the binary order defined as follows. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be the frame of discernment. For any $A \subseteq \Omega$, let $u(A)$ be the binary number $u_n u_{n-1} \ldots u_2 u_1$ 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 $A_1, \ldots, A_{2^n}$ to be indexed in such a way that $r(A_i) = i$ for all i, the vector representation of a mass function m on Ω is the column vector $\mathbf{m} = (m_1, \ldots, m_{2^n})'$, 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	$\{\omega_2, \omega_3\}$	$m(\{\omega_2, \omega_2\})$
8	111	$\{\omega_1, \omega_2, \omega_3\}$	$m(\{\omega_1, \omega_2, \omega_3\})$

Arranging the belief and plausibility numbers in vectors

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

and

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

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,

$$\mathbf{Bel} = \mathbf{BfrM} \cdot \mathbf{m}, \tag{2.10}$$

where **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 **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 Ω,

$$Bel(A \cup B) = \sum_{\omega \in A \cup B} m(\{\omega\}) = \sum_{\omega \in A} m(\{\omega\}) + \sum_{\omega \in B} m(\{\omega\}) = Bel(A) + Bel(B). \tag{2.11}$$

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 results 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 $A_1 \subset \ldots \subset A_r$. 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

$$\begin{aligned} Bel(A \cap B) &= \sum_{i=1}^{\min(i_1,i_2)} m(A_i) && \text{(2.12a)} \\ &= \min\left(\sum_{i=1}^{i_1} m(A_i), \sum_{i=1}^{i_2} m(A_i)\right) && \text{(2.12b)} \\ &= \min(Bel(A), Bel(B)). && \text{(2.12c)} \end{aligned}$$

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

$$\begin{aligned} Pl(A \cup B) &= 1 - Bel(\overline{A \cup B}) && \text{(2.13a)} \\ &= 1 - Bel(\overline{A} \cap \overline{B}) && \text{(2.13b)} \\ &= 1 - \min(Bel(\overline{A}), Bel(\overline{B})) && \text{(2.13c)} \\ &= \max(1 - Bel(\overline{A}), 1 - Bel(\overline{B})) && \text{(2.13d)} \\ &= \max(Pl(A), Pl(B)). && \text{(2.13e)} \end{aligned}$$

Properties (2.12c) and (2.13) characterize, respectively, *possibility* and *necessity* measures, which form the basis of Possibility theory introduced by Zadeh in [82]. 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 $pl : \Omega \to [0, 1]$ be the *contour* function of m, defined by $pl(\omega) = Pl(\{\omega\})$, for all $\omega \in \Omega$. For all $A \subseteq \Omega$,

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

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 Π is always a plausibility function for some consonant mass function, which can be recovered from π as explained in the following theorem [24].

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

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

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

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

$$m(\Omega) = \pi(\omega_n). \tag{2.15b}$$

Proof: Let A be a non empty subset of Ω. Let Π be the possibility measure induced by π and let Pl be the plausibility measure induced by m, given by (2.15). The possibility of A is

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

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,

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

$$= pl(\omega_{i_0}) - pl(\omega_{i_0+1}) + pl(\omega_{i0+1}) - \ldots - pl(\omega_n) \tag{2.17b}$$

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

$$= \Pi(A), \tag{2.17d}$$

which completes the proof. □

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

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

The corresponding mass function is

$$m(\{c\}) = 1 - 0.7 = 0.3 \tag{2.18a}$$
$$m(\{c,d\}) = 0.7 - 0.5 = 0.2 \tag{2.18b}$$
$$m(\{c,d,b\}) = 0.5 - 0.3 = 0.2 \tag{2.18c}$$
$$m(\{c,d,b,a\}) = 0.3. \tag{2.18d}$$

Possibility theory has a strong connection with the theory of Fuzzy Sets [81]. More precisely, if we receive evidence of the form "$\boldsymbol{\omega}$ is F", where F is a fuzzy subset of Ω with membership function μ_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 Ω. Its lower and upper envelopes are set functions defined as follows:

$$P_*(A) = \inf_{P \in \mathcal{P}} P(A), \tag{2.19a}$$
$$P^*(A) = \sup_{P \in \mathcal{P}} P(A). \tag{2.19b}$$

for all subsets A of Ω. Function P_* and P^* are called, respectively, *coherent lower and upper probabilities* [75]. Clearly,

$$P^*(A) = 1 - P_*(\overline{A}) \tag{2.20}$$

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 Ω. Because of the relation $Bel(A) = 1 - Pl(\overline{A})$, we also have $P(A) \leq Pl(A)$ for all A, or

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

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

$$\alpha : \Omega \times 2^\Omega \setminus \{\emptyset\} \to [0,1] \tag{2.22}$$

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

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

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

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

It can be shown [15] 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 [75, 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 $H_1 = \{(H,H),(H,T)\}$, $H_2 = \{(H,H),(T,H)\}$, and let $\mathcal{P}$ be the set of probability measures on Ω 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 H_1 may never lead to H_2). 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

$$P_*(H_1 \cup H_2) < P_*(H_1) + P_*(H_2) - P_*(H_1 \cap H_2), \tag{2.25}$$

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.

Exercices

1. Let Bel be a belief function on Ω 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 possess?

3. Represent the uncertainty about the outcome of the Ellsberg's experiment described in Section 1.3.4, 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 π be the following possibility distribution on $\Omega = \{a, b, c, d, e, f\}$:

ω	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 Ω and let Bel and Pl be the corresponding belief and plausibility functions. Show that, for any subset A of Ω, $Bel(A) > 0 \Rightarrow Pl(A) = 1$.

Chapter 3

Combination 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 $\Omega = \{\text{Peter}, \text{John}, \text{Mary}\}$. 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 : $m_2(\{\text{John}, \text{Mary}\}) = 0.6$, $m_2(\Omega) = 0.4$.

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 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?

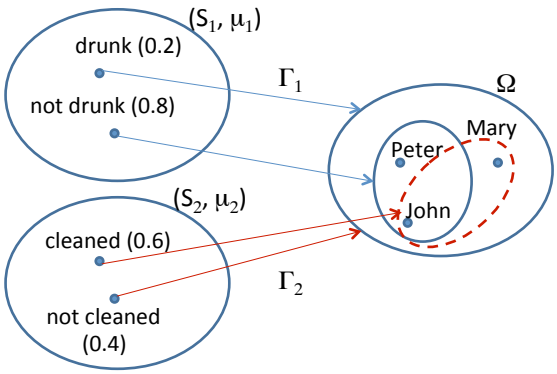


Figure 3.1: Combination of evidence in the murder example.

To answer this question, we need to describe the dependence between the two pieces of evidence by specifying a joint probability measure μ_{12} on the product space $S_1 \times S_2$. Independence between the two pieces of evidence corresponds to the case where μ_{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	$\{\text{John}, \text{Mary}\}$, 0.12	Ω, 0.08
not drunk	$\{\text{John}\}$, 0.48	$\{\text{Peter}, \text{John}\}$, 0.32

We then get the following combined mass function,

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

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	Ω, 0.08
not drunk	∅, 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	Ω, 0.08/0.52
not drunk	∅, 0	{Peter, John}, 0.32/0.52

which yields the following combined mass function,

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

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 $\{(s_1, s_2) \in S_1 \times S_2 | \Gamma_1(s_1) \cap \Gamma_2(s_2) \neq \emptyset\}$. 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 Ω. Dempster's rule is the partial binary operation $\oplus$ on $\mathcal{M}$ defined by

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

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

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

for all $A \subseteq \Omega$. It can be shown (see [58, Chapter 2]) that m, Bel and Pl can be uniquely recovered from Q using the following equations:

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

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

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

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

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

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

for all $A \subseteq \Omega$.

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

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

or, equivalently,

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

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*

$$(Q_1 \oplus Q_2)(A) = K Q_1(A) \cdot Q_2(A), \tag{3.12}$$

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

Proof. For any subset A of Ω, 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 $m_1, \ldots, m_n$. 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

$$Q(A) = K Q_1(A) \ldots Q_n(A) \tag{3.13}$$

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:

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

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 $m_1, \ldots, m_n$ at once using the following formula, which extends (3.3):

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

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

$$(m_1 \oplus \ldots \oplus m_n)(\emptyset) = 0, \tag{3.15b}$$

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

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

As mentioned above, κ 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* [58], defined as

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

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.,

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

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 Ω, in view, e.g., of selecting the most plausible element (see Section **??**). Then, by noticing that $pl(\omega) = Q(\{\omega\})$ for any $\omega \in \Omega$, we can obtain the contour function pl of $m_1 \oplus \ldots \oplus m_n$ as the product of the contour functions of the m_i's:

$$pl = pl_1 \ldots pl_n, \tag{3.19}$$

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 $\boldsymbol{\omega} \in B$ for some $B \subseteq \Omega$ such that $P(B) \neq 0$. The conditional probability measure is defined as

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

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 focussed on B:

$$m(\cdot|B) = m \oplus m_B. \tag{3.21}$$

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

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

where the normalizing constant K is

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

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

$$\begin{aligned} Pl(A|B) &= \sum_{C : C \cap A \neq \emptyset} m(C|B) && \text{(3.23a)} \\ &= Pl(B)^{-1} \sum_{C : C \cap A \neq \emptyset} \sum_{D : D \cap B = C} m(D) && \text{(3.23b)} \\ &= Pl(B)^{-1} \sum_{D : D \cap B \cap A \neq \emptyset} m(D) && \text{(3.23c)} \\ &= \frac{Pl(A \cap B)}{Pl(B)}. && \text{(3.23d)} \end{aligned}$$

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

$$\begin{aligned} Bel(A|B) &= 1 - Pl(\overline{A}|B) && \text{(3.24a)} \\ &= 1 - \frac{Pl(\overline{A} \cap B)}{Pl(B)} && \text{(3.24b)} \\ &= 1 - \frac{1 - Bel(A \cup \overline{B})}{1 - Bel(\overline{B})} && \text{(3.24c)} \\ &= \frac{Bel(A \cup \overline{B}) - Bel(\overline{B})}{1 - Bel(\overline{B})}. && \text{(3.24d)} \end{aligned}$$

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 [78]. 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 tranform [38], 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^Ω.

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 $\overline{\{\omega\}}$ can be performed in time proportional to the size of the frame [6].

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 Ω 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 [43, 72, 7, 31, 20]. The simplest, yet quite effective approach, is the Summarization algorithm [43], which works as follows. Let $F_1, \ldots, F_n$ be the focal sets of m ranked by decreasing mass, i.e., $m(F_1) \geq m(F_2) \geq \ldots \geq m(F_n)$. If n exceeds some the maximum allowed number k of focal sets, then the $n - k$

focal sets F_i, $i = k+1, \ldots, n$ with the smallest masses are replaced by their union, and m is approximated by the mass function m' defined as

$$m'(F_i) = m(F_i), \quad i = 1, \ldots, k, \tag{3.25a}$$

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

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., [49]).

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 μ 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 $\boldsymbol{\omega}$. 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* [69]. In that case, assuming the codes s_1 and s_2 to be used, we can deduce that $\boldsymbol{\omega} \in \Gamma_\cup(s_1, s_2) = \Gamma_1(s_1) \cup \Gamma_2(s_2)$. This assumption results in the following binary operation, called the *disjunctive rule of combination* [25, 66]:

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

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

The resulting mass function is

$$(m_1 \cup m_2)(\{a,b\}) = 0.05 + 0.2 + 0.02 + 0.08 = 0.35 \tag{3.27a}$$
$$(m_1 \cup m_2)(\{b,c\}) = 0.25 \tag{3.27b}$$
$$(m_1 \cup m_2)(\{a,c\}) = 0.03 + 0.15 = 0.18 \tag{3.27c}$$
$$(m_1 \cup m_2)(\Omega) = 0.1 + 0.12 = 0.22. \tag{3.27d}$$

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

$$(Bel_1 \cup Bel_2)(A) = Bel_1(A)Bel_2(A), \tag{3.28}$$

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 [25]. It is defined as follows:

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

for all $A \subseteq \Omega$, $A \neq \emptyset$, and $(m_1 \star m_2)(\emptyset) = 0$. 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

$$(m_1 \uplus m_2)(\{a,b\}) = 0.05 + 0.08 = 0.13 \tag{3.30a}$$
$$(m_1 \uplus m_2)(\{b\}) = 0.2 \tag{3.30b}$$
$$(m_1 \uplus m_2)(\{b,c\}) = 0.25 \tag{3.30c}$$
$$(m_1 \uplus m_2)(\{a\}) = 0.02 + 0.03 + 0.12 = 0.17 \tag{3.30d}$$
$$(m_1 \uplus m_2)(\{c\}) = 0.15 \tag{3.30e}$$
$$(m_1 \uplus m_2)(\Omega) = 0.1. \tag{3.30f}$$

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 μ_{12} on $S_1 \times S_2$, with marginals μ_1 and μ_2. To each joint measure μ_{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 μ_{12}^*, among all joints measures with marginals μ_1 and μ_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 $\boldsymbol{\omega} \in A$ 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 [58] 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 ∞ 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

$$(m_1 \oplus m_2)(A) = 1 - (1-p_1)(1-p_2) \tag{3.31}$$
$$(m_1 \oplus m_2)(\Omega) = (1-p_1)(1-p_2). \tag{3.32}$$

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

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

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

$$A^{w_1} \oplus A^{w_2} = A^{w_1+w_2}. \tag{3.34}$$

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

$$m = \bigoplus_{i=1}^{n} A_i^{w_i}, \tag{3.35}$$

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 [58], a particular "canonical" decomposition can be obtained as follows:

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

with

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

where $\mathcal{C}$ is the core of m. Shafer [58] called function $w : 2^\Omega \to [0, +\infty]$ defined by (3.37) the *assessment of evidence* associated with m. As shown in [58], a function $w : 2^\Omega \to [0, +\infty]$ is an assessment of evidence (for some separable mass function m) if and only of $w(\emptyset) = 0$, w assigns the value $+\infty$ to exactly one subset $\mathcal{C}$ of Ω, and w assigns the value zero to every subset of Ω not contained in $\mathcal{C}$. 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 $\mathcal{C} = \Omega$, then (3.37) has the following simpler expression,

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

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 [21] .

Proposition 3.1 *Let m be a mass function with focal sets $A_1, \ldots, A_r$ 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*

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

$$w(B) = +\infty \tag{3.39b}$$

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

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

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

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

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

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

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

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

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

$$w(A_r) = +\infty \tag{3.41b}$$

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

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

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

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

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

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

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

$$= \frac{\pi_k - \pi_{k+1}}{\pi_k} \pi_k \tag{3.43c}$$

$$= \pi_k - \pi_{k+1} \tag{3.43d}$$

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

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

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

Exercices

1. 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.

2. Let m be a mass function on Ω and B a non-empty subset of Ω.

   (a) Express the conditional belief function $Bel(.|B)$ as a function of Bel.

   (b) What does this formula become when Bel is a probability measure?

3. 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.

Chapter 4

Least commitment principle

In many situations, a belief functions 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 [40]. 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 [34]. 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. Applications of the Least Commitment Principles (LCP) will then be given in Section 4.3.

4.1 Inclusion relations

Let m_1 and m_2 be two mass functions on the same frame Ω. 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

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

and

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

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 [25]. This relation will be denoted by $\sqsubseteq_{Bel}$:

$$m_1 \sqsubseteq_{Bel} m_2 \text{ iff } Bel_1 \geq Bel_2. \tag{4.3}$$

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

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

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 Ω. Its greatest element is the vacuous mass function $m_?$, which assigns null degrees of belief to all proper subsets of Ω. 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

$$m_1 \sqsubseteq_{Bel} m_2 \Leftrightarrow pl_1 \leq pl_2. \tag{4.5}$$

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

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

and

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

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 [25]. 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,

$$m_1 \sqsubseteq_Q m_2 \Leftrightarrow pl_1 \leq pl_2. \tag{4.8}$$

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* $m_2 \sqsubseteq_Q m_1$.

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 [80, 39, 25]. 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:

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

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:

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

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* [68].

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

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

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

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

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

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

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

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

The specialization matrix is thus

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

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:

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

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

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

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

The corresponding generalization matrix is

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

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:

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

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 Ω 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 [21], where it was even extended to non-separable mass function, using the general canonical decomposition introduce by Smets in [67]. 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 $m_1 \oplus m_2 \sqsubseteq_w m_2$. Furtermore, 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 [34], the situation is less clear in belief function theory, in which several notions coexist. We will review some of the main definitions with there 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 $\boldsymbol{\omega} \in A$ for some non-empty subset A of Ω. 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

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

be such a measure. 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 $h(2) = 1$.

H_2 is quite natural and H_3 is a matter of convention. The only non-trivial axiom is H_1; it has the following meaning. Consider a partition of Ω into r subsets of s elements. Characterizing an element of Ω 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 $H3$.

It can be shown [40] that the only function H verifying these three axioms is defined by

$$h(n) = \log_2 n. \tag{4.18}$$

The function $H : 2^\Omega \setminus \emptyset \to \mathbb{R}_+$ 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:

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

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

$$m_1 \sqsubseteq m_2 \Rightarrow N(m_1) \leq N(m_2). \tag{4.20}$$

The nonspecificity was shown by Ramer [54] to be the only function satisfying the five axioms below. 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[1] of $m^{\Omega\times\Theta}$ on Ω and Θ are the mass functions obtained by transferring each mass $m^{\Omega\times\Theta}(C)$ to the projections of C on Ω and Θ, respectively. The marginals m^Ω and m^Θ are said to be non-interactive if, for all $C \subseteq \Omega \times \Theta$,

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

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^Ω and m^Θ, then

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

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

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

[1]The notion of marginal mass function will be further studied in Chapter 5.

(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 Ω that have equal cardinalities.

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

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

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

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

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

$$m(C) = m_1(C) = m_2(C_1), etc.$$

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 disjoints focal sets. This aspect may be described using other measures that extend the Shannon entropy.

4.2.2 Entropy-like measures

Seeing a mass function as the density of a random set [51], we could define its entropy as

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

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 the measure of dissonance defined

by

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

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

The term

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

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

$$-\log_2[1-K(A)] \tag{4.27}$$

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

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

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, defined by the formula:

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

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

4.2.3 Other uncertainty measures

Other uncertainty measures have been proposed. For example, the total uncertainty in a belief function may be defined the aggregate uncertainty,

defined as the maximum value of the Shannon entropy over all compatible probabilities:

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

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 inteactive ones. However, there is no proof that it is the only measure verifying all these properties.

A different approach was proposed by Smets [65], 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

$$I(m_1 \oplus m_2) = I(m_1) + I(m_2) \tag{4.31}$$

as well as some other trivial requirements lead to

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

where $c(A)$ is a constant depending on A. Choosing $c(A) = 1$, we get

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

It is obvious that

$$m_1 \sqsubseteq_Q m_2 \Rightarrow I(m_1) \geq I(m_2). \tag{4.34}$$

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 two 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 Ω. We may do it by asking him to give a strictly increasing sequence

$$A_1 \subset A_2 \subset \ldots \subset A_n$$

of subsets of Ω with their degrees of belief

$$0 < \alpha_1 < \alpha_2 < \ldots < \alpha_n = 1.$$

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 α_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 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 Ω, 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 will 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 $C \subseteq \overline{A}$:

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

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

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

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

4.3.3 Partial beliefs specifications

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

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

under the constraints:

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

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

$$m(\emptyset) = 0. \tag{4.37d}$$

4.3.4 Combination of mass functions with unknown dependence

Let us consider two sources (U_1, μ_1, Γ_1) and (U_2, μ_2, Γ_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 μ_{12} on $U_1 \times U_2$, whose marginals are μ_1 and μ_2.

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

$$p_{ij} = \mu(\{(u, v) \in U_1 \times U_2 | \Gamma_1(u) = A_i, \Gamma_2(v) = B_j\}). \tag{4.38}$$

Assuming both sources to be reliable, the combined mass function m has the following expression

$$m(A) = \sum_{A_i \cap B_j = A} p_{ij}. \tag{4.39}$$

When the dependence between the two sources is unknown, the p_{ij}'s are unknown, but we can find the least committed mass function of the form (4.39) by solving the following linear optimization problem:

$$\max_{p_{ij}} \sum_{i,j} p_{ij} \log_2 |A_i \cap B_j| \tag{4.40a}$$

under the constraints:

$$\sum_{i,j} p_{ij} = 1 \tag{4.40b}$$

$$\sum_{i} p_{ij} = q_j, \quad j = 1, \ldots, s \tag{4.40c}$$

$$\sum_{j} p_{ij} = p_i, \quad i = 1, \ldots, r. \tag{4.40d}$$

Chapter 5

Reasoning with multiple frames

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 [79]. 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 [63]. In evidential reasoning with uncertain information, we thus have to express belief functions in different frames with varying granularities.

5.1 Refinement and coarsening

Let Ω and Θ be two frames of discernment. We say that Ω is a *refinement* of Θ (or, equivalently, Θ is a *coarsening* of Ω) if elements of Ω can be obtained by splitting some or all of the elements of Θ (Figure 5.1) [58]. Formally, Ω is a refinement of a frame Θ iff there is a mapping $\rho : 2^\Theta \to 2^\Omega$ (called a *refining*) such that:

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

Let m^Θ be a mass function representing some piece of evidence expressed in the frame Θ, and let Ω be a refinement of Θ. We can carry m^Θ from Θ

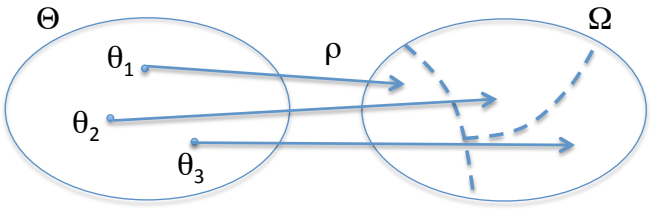


Figure 5.1: Refinement of a frame of discernment.

to Ω 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^Θ in Ω: for all $B \subseteq \Omega$,

$$m^{\Theta\uparrow\Omega}(B) = \begin{cases} m^\Theta(A) & \text{if } B = \rho(A), A \subseteq \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

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

$$\underline{\rho}^{-1}(B) = \{\theta \in \Theta | \rho(\{\theta\}) \subseteq B\}, \tag{5.2a}$$

$$\overline{\rho}^{-1}(B) = \{\theta \in \Theta | \rho(\{\theta\}) \cap B \neq \emptyset\}, \tag{5.2b}$$

for any subset B of Ω. The subsets $\underline{\rho}^{-1}(B)$ and $\overline{\rho}^{-1}(B)$ are called, respectively, the *inner reduction* and the *outer reduction* of B [58]. When computing the image by ρ of the inner or the outer reduction of B, we do not recover B in general. The following relations hold:

$$\rho[\underline{\rho}^{-1}(B)] \subseteq B \subseteq \rho[\overline{\rho}^{-1}(B)], \tag{5.3}$$

and the inclusions may be strict.

Let us now assume that we have a mass function m^Ω defined in Ω. In principle, it can be carried to Θ 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^Ω in Θ: for all subset A of Θ,

$$m^{\Omega\downarrow\Theta}(A) = \sum_{\overline{\rho}^{-1}(B)=A} m^\Omega(B). \tag{5.4}$$

We may observe that, in the process of carrying m^Ω from Ω to the coarser frame Θ, some information may be lost. In particular, if $m^{\Omega\downarrow\Theta}$ is carried back to Ω, we will not recover m^Ω in general. The resulting mass function will usually be strictly less informative than m^Ω according to the strong inclusion relation $\sqsubseteq$ (cf. Section 4.1.2), because any mass $m^\Omega(B)$ initially assigned to B is now assigned to a superset $\rho[\overline{\rho}^{-1}(B)]$. This actually makes sense: since some information is lost, we get a less informative mass function. This is the reason why the outer reduction $\overline{\rho}^{-1}$ is used in (5.4).

If two mass functions m^Ω and $m^{\Omega'}$

Example 5.1 *Consider, for instance, a ground detector and a sky detector.*

5.2 Special case of product spaces

5.2.1 Marginalization and vacuous extension

Let us now assume that we have two frames Ω_X and Ω_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 Ω_X and Ω_Y. For instance, we can define the following mapping ρ from 2^{Ω_X} to $2^{\Omega_X \times \Omega_Y}$:

$$\rho(A) = A \times \Omega_Y, \tag{5.5}$$

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

The vacuous extension of a mass function m^X from Ω_X to Ω_{XY} is obtained by transferring each mass $m^{\Omega_X}(B)$ for any subset B of Ω_X to the cylindrical extension of B (Figure 5.2):

$$m^{X\uparrow XY}(A) = \begin{cases} m^X(B) & \text{if } A = B \times \Omega_Y \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

Conversely, let m^{XY} be a joint mass function on the product space Ω_{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 Ω_X. We then have to compute the restriction of m^{XY} to the coarser frame Ω_X:

$$m^{XY\downarrow X}(A) = \sum_{B\downarrow\Omega_X = A} m^{XY}(B), \tag{5.7}$$

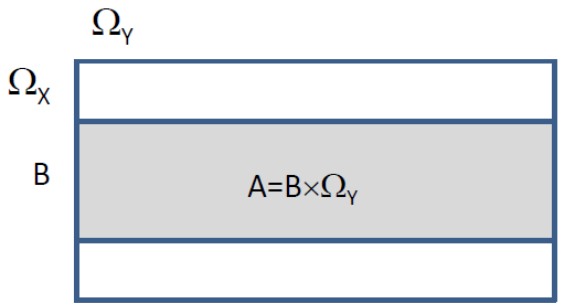


Figure 5.2: Vacuous extension.

where $B \downarrow \Omega_X$ denotes the *projection* of B on Ω_X (see Section 1.2.1). 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 functions is called *marginalization* (Figure 5.3).

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

$$m^{XY\downarrow X}(\{x\}) = \sum_{y \in \Omega_y} m^{XY}(\{x, y\}), \tag{5.8}$$

  which corresponds to probabilistic marginalization.

5.2.2 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:

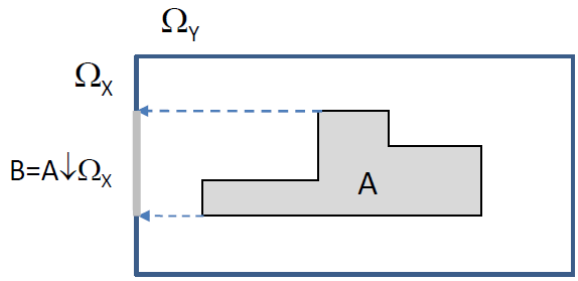


Figure 5.3: Marginalization.

- 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 Ω_{XY} and combining $m^{X\uparrow XY}$ with m^{XY}. The combined joint mass function can then be marginalized on Ω_Y. Formally,

$$m^Y = \left(m^{X\uparrow XY} \oplus m^{XY}\right)^{\downarrow Y}. \tag{5.9}$$

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, (5.9) can be written more compactly as

$$m^Y = \left(m^X \oplus m^{XY}\right)^{\downarrow Y}. \tag{5.10}$$

Equation (5.10) is clearly a generalization of (1.5).

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 [62, 3, 64].

In Section 5.4 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.

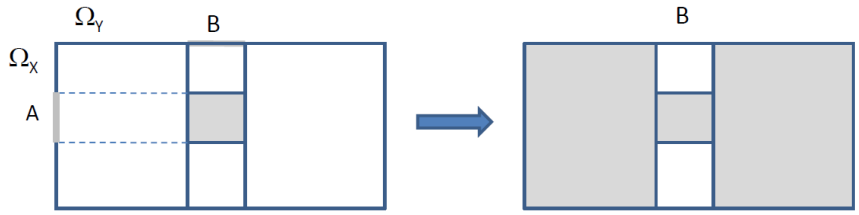


Figure 5.4: Conditional embedding operation. The mass on $m^X(A|B)$ is transferred to $(A \times \Omega_Y) \cup (\Omega_X \times \overline{B})$.

5.3 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 Ω, a subset B of Ω and m_B the logical mass function such as $m_B(B) = 1$, the conditional belief function $m(\cdot|B)$ is:

$$m(\cdot|B) = m \oplus m_B. \tag{5.11}$$

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 Ω whose core is included in B, its deconditioning yields the mass function m' defined by

$$m'(C) = \begin{cases} m(A) & \text{if } C = A \cup \overline{B} \text{ for some } A \subseteq B, \\ 0 & \text{otherwise,} \end{cases} \tag{5.12}$$

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 Ω_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:

$$m^X(\cdot|B) = \left(m^{XY} \oplus m_B^{Y\uparrow XY}\right)^{\downarrow X}. \tag{5.13}$$

Conversely, let $m^X(\cdot|B)$ represents your beliefs on Ω_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 (5.12). As shown in [66], the results is

$$m^{XY}(C) = \begin{cases} m^X(A|B) & \text{if } C = (A \times \Omega_Y) \cup (\Omega_X \times \overline{B}) \text{ for some } A \subseteq \Omega_X, \\ 0 & \text{otherwise.} \end{cases} \tag{5.14}$$

5.4 Applications

5.4.1 Discounting

Let us assume that we receive a mass function m_0^Ω from a source S, describing some information about ω. 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* [58, 66].

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 $\mathcal{R}$,

$$m^\mathcal{R}(\{R\}) = 1 - \alpha \tag{5.15a}$$

$$m^\mathcal{R}(\{NR\}) = \alpha, \tag{5.15b}$$

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

$$m^\Omega(\cdot|R) = m_0^\Omega. \tag{5.16}$$

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

$$m^\Omega(\Omega|NR) = 1. \tag{5.17}$$

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:

$$m_1^{\Omega\times\mathcal{R}}(\Omega \times \{R\}) = m^\mathcal{R}(\{R\}) = 1 - \alpha, \tag{5.18a}$$

$$m_1^{\Omega\times\mathcal{R}}(\Omega \times \{NR\}) = m^\mathcal{R}(\{NR\}) = \alpha. \tag{5.18b}$$

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

$$m_2^{\Omega\times\mathcal{R}}\left((A\times\mathcal{R})\cup(\Omega\times\{NR\})\right) = m_0^{\Omega}(A), \tag{5.19}$$

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

$${}^{\alpha}m^{\Omega}(A) = \begin{cases} (1-\alpha)m_0^{\Omega}(A) & \text{if } A \neq \Omega, \\ (1-\alpha)m_0^{\Omega}(\Omega) + \alpha & \text{if } A = \Omega. \end{cases} \tag{5.20}$$

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

$${}^{\alpha}m_0^{\Omega} = (1-\alpha)m_0^{\Omega} + \alpha\, m_?^{\Omega}. \tag{5.21}$$

The corresponding belief function has the following simple expression,

$${}^{\alpha}Bel^{\Omega}(A) = \begin{cases} (1-\alpha)Bel_0^{\Omega}(A) & \text{if } A \neq \Omega \\ 1 & \text{otherwise.} \end{cases} \tag{5.22}$$

5.4.2 Generalized Bayes' Theorem

The Generalized Bayes' Theorem (GBT) is due to Smets [66]. 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 Ω_X and Ω_Y. Let $\Omega_Y = \{y_1, \ldots, y_n\}$ and $m^X(\cdot|y_k) = m_k^X$, for $k = 1, \ldots, n$ 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 Ω_X is a set of symptoms and Ω_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 $k = 1, \ldots, n$ 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 Ω_X.

This problem can be solved by applying the following steps [66]:

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

$$m_k^{XY}(C \times \Omega_Y \cup \Omega_X \times \overline{\{y\}}) = m_k^X(C) \tag{5.23}$$

for all $C \subseteq \Omega_X$.

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

$$m_k^Y(\overline{\{y_k\}}|A) = 1 - Pl_k^X(A) \tag{5.24a}$$
$$m_k^Y(\Omega_Y|A) = Pl_k^X(A), \tag{5.24b}$$

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

$$m^Y(\cdot|A) = \bigoplus_{k=1}^{n} \overline{\{y_k\}}^{-\log Pl_k^{\mathbf{X}}(A)}. \tag{5.25}$$

Equation (5.25) 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_ℓ 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 $\overline{\{y_k\}}^{-\log Pl_k^{\mathbf{X}}(A)}$ is

$$\kappa = \prod_{k=1}^{n} \left(1 - Pl_k^{\mathbf{X}}(A)\right). \tag{5.26}$$

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

$$Q_k^Y(B|A) = \begin{cases} Pl_k^{\mathbf{X}}(A) & \text{if } y \in B, \\ 1 & \text{otherwise,} \end{cases} \tag{5.27}$$

for all $B \subseteq \Omega_Y$. Consequently, the combined commonality function corresponding to (5.25) is

$$Q^Y(B|A) = K \prod_{y_k \in B} Pl_k^{\mathbf{X}}(A), \tag{5.28}$$

where $K = (1-\kappa)^{-1}$. Smets [66] shows that the corresponding plausibility function is:

$$Pl^Y(B|A) = K \left[1 - \prod_{y_k \in B} (1 - Pl_k^X(A)) \right]. \tag{5.29}$$

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 (5.25) and combine it with the prior m_0^Y. The result is a Bayesian mass function defined as

$$m_1^Y(\{y_k\}|A) = \frac{m_k^X(A) m_0^Y(\{y_k\})}{\sum_{\ell=1}^n m_\ell^X(A) m_0^Y(\{y_\ell\})} \tag{5.30}$$

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 Ω_X given y need not be Bayesian.

The second remarkable property is related to the notion of *cognitive independence* [58]. Assume that we have three variables X, Y and Z. Variables X and Z are said to be *cognitively independent* conditionally on Y if, or all $A \subseteq \Omega_X$, $B \subseteq \Omega_Z$ and $y_k \in \Omega_Y$,

$$pl^{XZ}(A \times B|y_k) = pl^X(A|y_k) \cdot pl^Z(B|y_k). \tag{5.31}$$

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 Ω_Y given the facts $X \in A$ and $Z \in B$ directly using (5.25), 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:

$$m^Y(\cdot|X \in A, Z \in B) = m^Y(\cdot|X \in A) \oplus m^Y(\cdot|Z \in B). \tag{5.32}$$

Exercises

1. Let $\Theta = \{\theta_1, \theta_2\}$ and $\Omega = \{a, b, c, d\}$ be two frames of discernment, and let ρ be the refinement defined by $\rho(\{\theta_1\}) = \{a, b\}$ and $\rho(\{\theta_2\}) = \{c, d\}$.

   (a) Compute the vacuous extension on Ω 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 Θ 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 6

Belief functions on infinite spaces

Until now, the presentation of belief functions has been restricted to the case where Ω 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 [57, 59]. 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 [51], it will provide a solid foundation for a theory of belief functions in infinite spaces.

6.1 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).

6.1.1 Definitions

Let us consider a set Ω and an algebra $\mathcal{B}$ of subsets of Ω (see Section 1.3.1). A belief function on $\mathcal{B}$ is a function $Bel : \mathcal{B} \to [0,1]$ verifying the following three conditions:

1. $Bel(\emptyset) = 0$;
2. $Bel(\Omega) = 1$;
3. For any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,

$$Bel\left(\bigcup_{i=1}^{k} B_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} B_i\right). \tag{6.1}$$

Similarly, a plausibility function can be defined as a function $Pl : \mathcal{B} \to [0,1]$ such that:

1. $Pl(\emptyset) = 0$;
2. $Pl(\Omega) = 1$;
3. For any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,

$$Pl\left(\bigcap_{i=1}^{k} B_i\right) \leq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Pl\left(\bigcup_{i \in I} B_i\right). \tag{6.2}$$

It is clear that, whenever Bel is a belief function, Pl defined by $Pl(A) = 1 - Bel(\overline{A})$ is a plausibility function.

6.1.2 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 $\mathcal{B}$ [15]. More precisely, let S be a set, $\mathcal{A}$ an algebra of subsets of S, μ a finitely additive probability measure on $\mathcal{A}$, and $\Gamma : S \to 2^{\Omega}$ a multi-valued mapping. We can define two inverses of Γ:

1. The lower inverse

$$\Gamma_*(B) = B_* = \{s \in S | \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\}; \tag{6.3}$$

2. The upper inverse

$$\Gamma^*(B) = B^* = \{s \in S | \Gamma(s) \cap B \neq \emptyset\}, \tag{6.4}$$

for all $B \in \mathcal{B}$. We say that Γ 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

$$\begin{aligned} B_* &= \{s \in S | \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\} && (6.5a)\\ &= \{s \in S | \Gamma(s) \subseteq B\} \cap \{s \in S | \Gamma(s) \neq \emptyset\} && (6.5b)\\ &= \{s \in S | \Gamma(s) \cap \overline{B} = \emptyset\} \cap \{s \in S | \Gamma(s) \neq \emptyset\} && (6.5c)\\ &= \overline{B^*} \cap \Omega^*. && (6.5d) \end{aligned}$$

We then have the following important theorem [50].

Theorem 6.1 *Let $\mathcal{A}$ be an algebra of subsets of a set S, μ a finitely additive probability measure on $\mathcal{A}$, $\mathcal{B}$ an algebra of subsets of a set Ω, and Γ 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 $B \in \mathcal{B}$,*

$$\mu_*(B) = K\mu(B_*), \tag{6.6a}$$
$$\mu^*(B) = K\mu(B^*) = 1 - \mu_*(\overline{B}), \tag{6.6b}$$

where $K = [\mu(\Omega^*)]^{-1}$. Then, μ_* is a belief function and μ^* is the dual plausibility function.

Proof. First, we have

$$\begin{aligned} \Omega^* &= \{s \in S | \Gamma(s) \neq \emptyset\} && (6.7a)\\ &= \{s \in S | \Gamma(s) \cap B \neq \emptyset\} \cup \{s \in S | \Gamma(s) \cap B = \emptyset, \Gamma(s) \neq \emptyset\} && (6.7b)\\ &= B^* \cup (\overline{B})_*. && (6.7c) \end{aligned}$$

Since $B^* \cap (\overline{B})_* = \emptyset$, we can deduce that $\mu^*(B) = 1 - \mu_*(\overline{B})$. To prove that μ_* verifies (2.5), we can remark that

$$\Gamma_*\left(\bigcap_i B_i\right) = \bigcap_i \Gamma_*(B_i) \tag{6.8}$$

and

$$\Gamma^*\left(\bigcup_i B_i\right) \supseteq \bigcup_i \Gamma_*(B_i). \tag{6.9}$$

Consequently, for any k and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,

$$\sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} \mu_* \left(\bigcap_{i\in I} B_i\right) =$$

$$\sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} K\mu \left(\bigcap_{i\in I} \Gamma_*(B_i)\right) = K\mu\left(\bigcup_{i\in I} \Gamma_*(B_i)\right)$$

$$\leq K\mu\left[\Gamma_*\left(\bigcup_i B_i\right)\right] = \mu_*\left(\bigcup_i B_i\right). \tag{6.10}$$

Thus, to define a belief function on $(\Omega, \mathcal{B})$, it suffices to define a finitely additive probability μ on an algebra $\mathcal{A}$ of subsets of a set S and a strongly measurable mapping Γ 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 [57] 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 μ a finitely additive probability. In Shafer's contructive proof (too complex to be reproduced here), $S = 2^\Omega$, $\mathcal{A}$ is an algebra of subsets of 2^Ω, and Γ is the identity mapping. However, as well shall see in Section 6.2, a much simpler representation can often be found in practice.

As shown by Shafer [59], any belief function Bel on $(\Omega, \mathcal{B})$ can be extended to $(\Omega, 2^\Omega)$ as

$$\widetilde{Bel}(A) = \sup\{Bel(B) | B \in \mathcal{B}, B \subseteq A\}, \tag{6.11}$$

for all $A \subseteq \Omega$. A belief function Bel on a σ-algebra $\mathcal{B}$ is *continuous* if, for any decreasing sequence $B_1 \supset B_2 \supset B_3 \supset \ldots$ of elements of $\mathcal{B}$,

$$\lim_{i\to+\infty} Bel(B_i) = Bel\left(\bigcap_i B_i\right). \tag{6.12}$$

If the algebras $\mathcal{A}$ and $\mathcal{B}$ of Theorem 6.1 are σ-algebras, and if μ is a countably additive probability, then the belief function Bel on $\mathcal{B}$ induced by a strongly measurable mapping Γ is continuous. To see this, we can notice that, for any decreasing sequence $B_1 \supset B_2 \supset B_3 \supset \ldots$ of elements of $\mathcal{B}$,

$\Gamma_*(B_1) \supset \Gamma_*(B_2) \supset \Gamma_*(B_3) \supset \ldots$ is a decreasing sequence of elements of $\mathcal{A}$. Consequently,

$$\lim_{i\to+\infty} \mu_*(B_i) = \lim_{i\to+\infty} K\mu\left(\Gamma_*(B_i)\right) = K\mu\left(\bigcap_i \Gamma_*(B_i)\right) = K\mu\left[\Gamma_*\left(\bigcap_i B_i\right)\right] = \mu_*\left(\bigcap_i B_i\right). \quad (6.13)$$

6.1.3 Dempster's rule

Assume that we have n sources $(S_i, \mathcal{A}_i, \mu_i, \Gamma_i)$ for $i = 1, \ldots, n$, where each Γ_i is a multi-valued mapping from S_i to 2^Ω. Then, the combined source $(S, \mathcal{A}, \mu, \Gamma_\cap)$ can be defined as follows [15]:

$$S = S_1 \times S_2 \ldots \times S_n, \quad (6.14a)$$
$$\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2 \ldots \otimes \mathcal{A}_n, \quad (6.14b)$$
$$\mu = \mu_1 \times \mu_2 \ldots \times \mu_n, \quad (6.14c)$$
$$\Gamma_\cap(s) = \Gamma_1(s_1) \cap \Gamma_2(s_2) \cap \ldots \cap \Gamma_n(s_n), \quad (6.14d)$$

where $\mathcal{A}$ is the algebra generated by subsets of S of the form $A_1 \times \ldots \times A_n$, with $A_i \in \mathcal{A}_i$, and μ is the product measure. The belief function Bel induced by $(S, \mathcal{A}, \mu, \Gamma_\cap)$ can then be written as $Bel_1 \oplus \ldots \oplus Bel_n$, 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 $\Gamma_\cap(s) \neq \emptyset$,

$$Bel(B) = \frac{\mu\left(\{s \in S | \Gamma_\cap(s) \neq \emptyset, \Gamma_\cap(s) \subseteq B\}\right)}{\mu(\{s \in S | \Gamma_\cap(s) \neq \emptyset\})}, \quad (6.15)$$

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 (6.15).

This combination rule, introduced by Dempster in Ref. [15], clearly generalizes that introduced in Chapter 3 for the finite case. The consideration of the product probability measure in (6.14c) 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

$$Q(B) = K\mu\left(\{s \in S | \Gamma(s) \supseteq B\}\right), \quad (6.16)$$

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 $(S, \mathcal{A}, \mu, \Gamma_\cap)$, andw e can easily check that

$$Q(B) = \frac{1}{1-\kappa} \prod_{i=1}^{n} Q_i. \tag{6.17}$$

6.2 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.

6.2.1 Consonant random closed sets

Let us assume that $\Omega = \mathbb{R}^d$ and $\mathcal{B} = 2^\Omega$. Let π be an upper semi-continuous map from $\mathbb{R}^d$ to $[0,1]$, i.e., for any $s \in [0,1]$, the set

$$^s\pi = \{x \in \mathbb{R}^d | \pi(x) \geq s\} \tag{6.18}$$

is closed. Furthermore, assume that $\pi(x) = 1$ for some x. Let $S = [0,1]$, $\mathcal{A}$ be the Borel σ-field on $[0,1]$, μ the Lebesgue measure, and Γ the mapping defined by $\Gamma(s) = {}^s\pi$. Then Γ is strongly measurable and it defines a random closed set [51]. 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 $B \subseteq \mathbb{R}^d$:

$$\begin{aligned} Pl(B) &= \mu(\{s \in S | {}^s\pi \cap B \neq \emptyset\}) && \text{(6.19a)} \\ &= \mu(\{s \in S | \exists x \in B, \pi(x) \geq s\}) && \text{(6.19b)} \\ &= \mu(\{s \in S | s \leq \sup_{x \in B} \pi(x)\}) && \text{(6.19c)} \\ &= \sup_{x \in B} \pi(x), && \text{(6.19d)} \end{aligned}$$

$$Bel(B) = 1 - Pl(\overline{B}) = 1 - \sup_{x \notin B} \pi(x) = \inf_{x \notin B} (1 - \pi(x)) \tag{6.20}$$

and

$$\begin{aligned} Q(B) &= \mu(\{s \in S | {}^s\pi \supseteq B\}) && \text{(6.21a)} \\ &= \mu(\{s \in S | \forall x \in B, \pi(x) \geq s\}) && \text{(6.21b)} \\ &= \mu(\{s \in S | s \leq \inf_{x \in B} \pi(x)\}) && \text{(6.21c)} \\ &= \inf_{x \in B} \pi(x). && \text{(6.21d)} \end{aligned}$$

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 8.3.

6.2.2 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 [18].

Definition and properties

Let (U, V) be a bi-dimensional random vector from $(S, \mathcal{A}, \mu)$ to $\mathbb{R}^2$ such that

$$\mu(\{s \in S | U(s) \leq V(s)\}) = 1. \tag{6.22}$$

The mapping

$$\Gamma : s \to \Gamma(s) = [U(s), V(s)], \tag{6.23}$$

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:

$$Bel((-\infty, x]) = \mu([U, V] \subseteq (-\infty, x]) = \mu(V \leq x) = F_V(x), \tag{6.24}$$

where F_V is the cumulative distribution function (cdf) of V, and

$$Pl((-\infty, x]) = \mu([U, V] \cap (-\infty, x] \neq \emptyset) = \mu(U \leq x) = F_U(x). \tag{6.25}$$

These functions are called, respectively, the lower and upper cdf of $[U, V]$. Now, for any $a \leq b$, we have

$$Bel([a, b]) = \mu([U, V] \subseteq [a, b]) = \mu(U \geq a; V \leq b), \tag{6.26}$$

$$\begin{aligned} Pl([a, b]) = \mu([U, V] \cap [a, b] \neq \emptyset) = \\ 1 - \mu([U, V] \cap [a, b] = \emptyset) = 1 - \mu(V < a) - \mu(U > b) \end{aligned} \tag{6.27}$$

and

$$Q([a,b]) = \mu([U,V] \supseteq [a,b]) = \mu(U \leq a; V \geq b). \tag{6.28}$$

We can observe that If $[U,V]$ is continuous, these probabilities can be computed by integrating the joint density $f(u,v)$. For instance,

$$Q([a,b]) = \int_{-\infty}^{a} \int_{b}^{+\infty} f(u,v) dv du, \tag{6.29}$$

$$Bel([a,b]) = \int_{a}^{b} \int_{u}^{b} f(u,v) dv du. \tag{6.30}$$

Conversely,

$$f(a,b) = -\frac{\partial^2 Q([a,b])}{\partial a \partial b} = -\frac{\partial^2 Bel([a,b])}{\partial a \partial b}. \tag{6.31}$$

Lower and upper quantiles

If the random vector (U,V) is continuous, we can define its lower and upper quantiles at level α, for any $\alpha \in (0,1)$, as:

$$q_*(\alpha) = F_U^{-1}(\alpha), \tag{6.32a}$$

$$q^*(\alpha) = F_V^{-1}(\alpha). \tag{6.32b}$$

By definition, $q_*(\alpha)$ and $q^*(\alpha)$ are thus, respectively, the values such that

$$Pl((-\infty, q_*(\alpha)]) = \alpha \tag{6.33a}$$

and

$$Bel((-\infty, q^*(\alpha)]) = \alpha \tag{6.33b}$$

or, equivalently,

$$Pl((q^*(\alpha), +\infty)) = 1 - \alpha. \tag{6.33c}$$

For any $\alpha \in (0, 0.5]$, we may compute the α-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 $Pl_x^{\mathbf{Y}}$, we may conclude that

$$Pl(\overline{(q_*(\alpha), q^*(1-\alpha)]}) \leq 2\alpha. \tag{6.34a}$$

or, equivalently,

$$Bel((q_*(\alpha), q^*(1-\alpha)]) \geq 1 - 2\alpha. \tag{6.34b}$$

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:

$$(Q_1 \oplus Q_2)([a,b]) = \frac{1}{1-\kappa} Q_1([a,b]) Q_2([a,b]), \tag{6.35}$$

where

$$\kappa = \mu([U_1, V_2] \cap [U_2, V_2] = \emptyset) \tag{6.36}$$

is the degree of conflict between the two random sets. To see this, we may observe that

$$(Q_1 \oplus Q_2)([a,b]) = \mu([U_1, V_1] \cap [U_2, V_2] \supseteq [a,b] | [U_1, V_1] \cap [U_2, V_2] \neq \emptyset), \tag{6.37}$$

which may be computed as

$$(Q_1 \oplus Q_2)([a,b]) = \frac{\mu([U_1, V_1] \supseteq [a,b], [U_2, V_2] \supseteq [a,b])}{\mu([U_1, V_1] \cap [U_2, V_2] \neq \emptyset)} \tag{6.38a}$$

$$= \frac{Q_1([a,b]) Q_2([a,b])}{1-\kappa}. \tag{6.38b}$$

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 [4]. 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
$\widehat{(Pl_1 \oplus Pl_2)}(A) = \frac{k}{N}$

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 $\Omega = \{\omega_1, \ldots, \omega_n\}$ the set of states of nature and by $\mathcal{C} = \{c_1, \ldots, c_r\}$ the set of consequences (or outcomes), an act can thus be formalized as a mapping f from Ω to $\mathcal{C}$. In most of this chapter, the three sets Ω, $\mathcal{C}$ and $\mathcal{F}$ will be assumed to be finite.

The desirability of the consequences can often be modeled by a *utility function* $u : \mathcal{C} \to \mathbb{R}$, 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 (ω_1)	Poor Economic Conditions (ω_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 $u[f_i(\omega_i)]$. 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 [52].

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 (ω_1) or poor (ω_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 is know, then the desirability of an act can be deduced from that of its consequences $f(\omega)$. 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 $\succcurlyeq$ on $\mathcal{F}$. This relation is interpreted as follows: given two acts f and g, $f \succcurlyeq g$ 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 $f \succcurlyeq g$ and not$(g \succcurlyeq f)$ (meaning that f is strictly more desirable than g) and an indifference relation $f \sim g$ iff $f \succcurlyeq g$ and $g \succcurlyeq f$ (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. The main purpose of this chapter is to review the main solutions, when uncertainty is described by belief functions. before that, we will start

Different decision problems arise, depending on the nature of the available information. 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 Ω are given, which gives rise to the expected utility (EU) model (Section 7.2.2). Final several axiomatic justification of the EU model will be reviewed in Section **??**.

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. A 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 [71] 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 mots 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., [44, 71]).

1. The *Laplace criterion* ranks acts according to the average utility of their consequences: $f_i \succeq f_k$ iff
$$\frac{1}{n}\sum_j u_{ij} \geq \frac{1}{n}\sum_j u_{kj}. \tag{7.1}$$

2. The *maximax criterion* considers, for each act, its more favorable consequence. We then have $f_i \succeq f_k$ iff
$$\max_j u_{ij} \geq \max_j u_{kj}. \tag{7.2}$$

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
$$\min_j u_{ij} \geq \min_j u_{kj}. \tag{7.3}$$

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

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

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

5. Finally, the *minimax regret criterion* considers an act f_i to me 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: we thus have $f_i \succeq f_k$ iff

$$\max_j (\max_\ell u_{\ell j} - u_{ij}) \leq \max_j (\max_\ell u_{\ell j} - u_{kj}). \qquad (7.5)$$

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 1.3.4). The maximax and maximin criteria correspond, respectively, to extreme optimistic and pessimistic (or conservative) attitudes of the DM. The Hurwicz criterions 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.

Each of the five criteria listed above induces a complete preorder over the set of acts. Let $\mathcal{F}^*$ be the set of greatest element (called the choice-set) for one of these preorders. Luce and Raiffa [44] have proposed a set of "postulates of rational choice" as a means of appraising the criteria. However, none of the five criteria satisfies all these postulates. For instance, it may be argued that $\mathcal{F}^*$ should be invariant under deleting a repetitious column, a requirement that is clearly violated by the Laplace criterion. Another compelling postulate is that "an act not belonging to $\mathcal{F}^*$ cannot be made to belong to $\mathcal{F}^*$ by adding new acts to $\mathcal{F}$. This postulate is violated by the

minimax regret criterion. Consider again, for instance, the data of Example 7.1. We have seen that $\mathcal{F}^* = \{f_1\}$ according to the minimax regret criterion. Now, consider a new act f_4 such that $u_{41} = 130,000$ and $u_{42} = -45,000$. With this new act, the regret of f_1 becomes 80,000 under ω_1 and 0 under ω_2: the maximal regret is thus 80,000. Now, the regret of f_2 becomes 30,000 under ω_1 and -45,000 under ω_2, with a maximal regret of 30,000. We can see that f_2 now becomes more desirable than f_1, as a consequence of adding a new act. As remarked by Szaniawski [71], this behavior can hardly be considered rational.

7.2.2 Decision-making with probabilities

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

$$EU(f_i) = \sum_j u_{ij} p_j. \tag{7.6}$$

According to the Maximum Expected Utility (MEU) principle, an act f_i is more desirable than an act f_k if its 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 70% chance that the economic situation will be poor. The expected utilities of acts f_1 and f_2 are*

$$EU(f_1) = 50,000 \times 0.4 + 30,000 \times 0.6 = 38,000 \tag{7.7a}$$
$$EU(f_2) = 100,000 \times 0.4 - 40,000 \times 0.6 = 16,000. \tag{7.7b}$$

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 [73]. We give hereafter a summary of their argument. Given a probability distribution on Ω, an act $f : \Omega \to \mathcal{C}$ 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 Morgentern argued that, to be rational, a preference relation should very 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 probability $\alpha \in [0, 1]$ such that

$$\alpha P + (1 - \alpha)R \sim Q, \tag{7.8}$$

   where $\alpha P + (1-\alpha)R$ is a compound lottery, which refers to the situation where you receive P with probability α and Q with probability $1 - \alpha$.

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

$$P \succeq Q \Leftrightarrow \alpha P + (1 - \alpha)R \succeq \alpha Q + (1 - \alpha)R. \tag{7.9}$$

We then have the following theorem.

Theorem 7.1 (Von Neumann and Morgentern) *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 $u : \mathcal{C} \to \mathbb{R}$ such that, for any two lotteries $P = (p_1, \ldots, p_r)$ and $Q = (q_1, \ldots, q_r)$,*

$$P \succeq Q \Leftrightarrow \sum_{i=1}^{r} p_i u(c_i) \geq \sum_{i=1}^{r} q_i u(c_i). \tag{7.10}$$

   Function u is unique up to an strictly increasing affine transformation.

7.2.3 Savage's theorem

Savage [56] proposed seven axioms that a preference relation among acts should verify. We will review the the first four, which are considered to be meaningful, the other three being mostly technical.

The first axiom (ordering) states that $\succcurlyeq$ 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 Ω.

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

$$(fEh)(\omega) = \begin{cases} f(\omega) & \text{if } \omega \in E \\ h(\omega) & \text{if } \omega \notin E. \end{cases} \tag{7.11}$$

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',

$$fEh \succcurlyeq gEh \Rightarrow fEh' \succcurlyeq gEh'. \tag{7.12}$$

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:

$$x \succcurlyeq y \Leftrightarrow (\forall f \in \mathcal{F},\ xEf \succcurlyeq yEf). \tag{7.13}$$

Lastly, the fourth axiom (Weak comparative ordering) can be stated as follows: for all events A and B and outcomes $x' \succ x$ and $y' \succ y$,

$$x'Ax \succcurlyeq xBx' \Rightarrow y'Ay \succcurlyeq yBy'. \tag{7.14}$$

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 $\succcurlyeq$ satisfies the above four axioms (plus three more technical ones) iff there is a finitely additive probability measure P and a function $u : \mathcal{C} \to \mathbb{R}$ such that f is at most preferable to g iff the expected utility of f is not smaller than that of g,

$$\forall f, g \in \mathcal{F}, \quad f \succcurlyeq g \Leftrightarrow \mathbb{E}_P[u \circ f] \geq \mathbb{E}_P[u \circ g]. \tag{7.15}$$

Moreover, P is unique and u is unique up to a positive affine transformation. Function u is called a *utility function*.

This is a very powerful result. It means that any DM verifying the above axioms

- Quantifies uncertainty about the states by a subjective probability measure;
- Ranks consequences according to a utility function that reflects his/her preferences;
- Evaluates acts according to the expected utility criterion.

7.3 Decision-making with belief functions

7.3.1 Upper and lower expected utility

7.3.2 Other approaches

7.3.3 Axiomatic justifications

Chapter 8

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 [14, 15, 17, 18]. 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 Θ, while the *prediction* problem is to construct a belief function about some new data Y to be generated from a random process depending on θ.

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 8.1). We will then present the two main methods of inference in the belief function framework: Dempster's method (Section 8.2) and the likelihood-based method (Section 8.3). A general prediction method will be then exposed in Section 8.4.

8.1 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.

8.1.1 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 [8].

Example 8.1 *Suppose X_1 and X_2 are iid with probability mass function*

$$\mathbb{P}_\theta(X_i = \theta - 1) = \mathbb{P}_\theta(X_i = \theta + 1) = \frac{1}{2}, \quad i = 1, 2, \tag{8.1}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Consider the following confidence set for θ,

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{otherwise.} \end{cases} \tag{8.2}$$

The corresponding confidence level $1 - \alpha = P_\theta(\theta \in C(X_1, X_2))$ is

$$\begin{aligned} 1 - \alpha = \mathbb{P}_\theta(\theta \in C(X_1, X_2) | X_1 \neq X_2)\mathbb{P}_\theta(X_1 \neq X_2) + \\ \mathbb{P}_\theta(\theta \in C(X_1, X_2) | X_1 = X_2)\mathbb{P}_\theta(X_1 = X_2) = \\ 1 \times 0.5 + (0.5)^2 = 0.75. \end{aligned} \tag{8.3}$$

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 θ 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. □

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 [8].

Example 8.2 *Suppose the distributions of a discrete random variable X for two values θ_1 and θ_2 of some parameter θ 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

$$p_1 = P_{\theta_1}(X \geq 2) = 0.04 + 0.037 + 0.033 = 0.11, \tag{8.4}$$

while the p-value against H_2 is

$$p_2 = P_{\theta_2}(X \geq 2) = 0.04 + 0.005 + 0.005 = 0.05. \tag{8.5}$$

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. □

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 8.1.3.

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.

8.1.2 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 θ, Bayes' rule allows us to compute the conditional pdf of θ given the data x, referred to as the *posterior distribution* of θ, and defined as

$$f(\theta|x) \propto \pi(\theta) f(x|\theta). \tag{8.6}$$

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 θ before observing the data. We should then select a "noninformative" prior distribution, but how can it be defined? If we can specificy a bounded support for θ, 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 1.3.4. 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 π 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 [35]. The Jeffreys prior is defined objectively as being proportional to the square root of the determinant of the Fisher information.

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \tag{8.7}$$

where the component (i, j) of the information matrix $I(\theta)_{ij}$ is

$$I(\theta)_{ij} = \mathbb{E}_\theta \left[\frac{\partial \log f_\theta(x)}{\partial \theta_i} \frac{\partial \log f_\theta(x)}{\partial \theta_j} \right]. \tag{8.8}$$

The motivation for this definition is that the Jeffreys prior is invariant under reparameterization: if φ is a one-to-one transformation and $\nu = \varphi(\theta)$, then the Jeffreys prior on ν 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 $X_1, \ldots, X_n$ from a Bernoulli distribution $\mathcal{B}(\theta)$. The Jeffreys prior on θ is the beta distribution $B(0.5, 0.5)$ whose pdf is displayed in Figure 8.1. We can see that extreme values of θ are considered a priori more probable that central values, which does represent non vacuous knowledge about θ.

8.1.3 Likelihood-based approach

Beside the frequentist and Bayesian schools of scientific inference, a third tradition can be traced back from Fisher's later work [30] to Barnard [5], Birnbaum [10] and Edward [27], 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 [10], "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 [10]. Let E denote a statistical model representing an experimental situation. Typically, E is composed of the parameter space Θ,

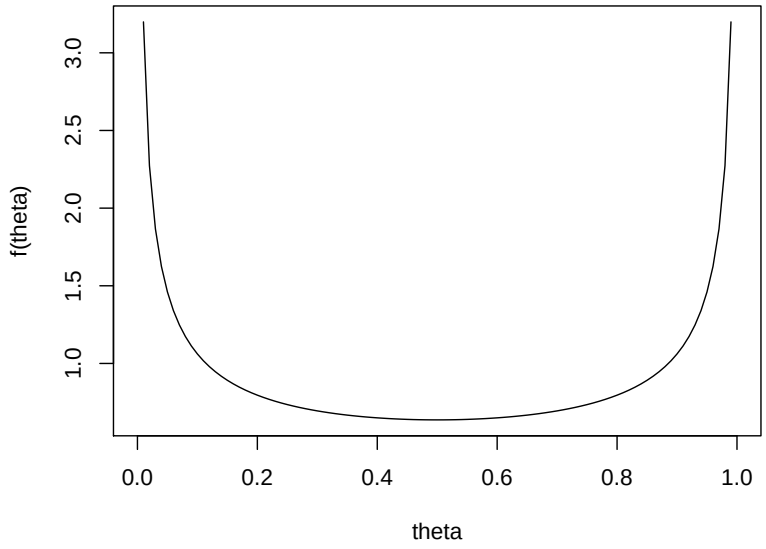


Figure 8.1: Jeffreys prior for a Bernoulli sample.

the sample space $\mathcal{X}$ and a probability mass or density function $f_\theta(x)$ for each $\theta \in \Theta$. Following Birnbaum [10], 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 Θ, 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 [10], 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 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 [30] and Barnard [5]. From a non Bayesian perspective, it was placed on firm ground by Birnbaum [10], 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 θ 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 [27, 2], repeatedly stressed that "probability and likelihood are quantities of an entirely different nature" [29] as, in particular, likelihoods are not additive. In Section 8.3, 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.

8.2 Dempster's method

The starting point of the theory of belief functions is a series of papers published by Dempster in 1960 [14, 15, 16, 17, 18], 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 [19]. Similar ideas are explored in [41, 42, 46] with a different terminology.

8.2.1 General method

The key idea underlying Dempster's method of inference is to model the data-generating mechanism by an equation

$$X = \varphi(\theta, W) \tag{8.9}$$

that relates the data X, the parameter θ and an unobserved auxiliary variable W with sample space $\mathcal{W}$ and known probability distribution $\mathbb{P}_W$ independent of θ, such that, for any measurable subset A of $\mathcal{X}$,

$$\mathbb{P}_\theta(X \in A) = \mathbb{P}_W(\varphi(\theta, W) \in A). \tag{8.10}$$

When X is a one-dimensional continuous random variable, a natural choice for W is $W = F_\theta(X)$, where F_θ is the cumulative distribution function (cdf) of X. The random variable W then has then a uniform distribution on $[0, 1]$ and (8.9) becomes

$$X = F_\theta^{-1}(W). \tag{8.11}$$

When W is discrete, (8.11) is still valid if F_θ^{-1} now denotes the generalized inverse of F_θ,

$$F_\theta^{-1}(w) = \inf\{x | F_\theta(x) \geq w\}. \tag{8.12}$$

From now on, we will assume that $\mathcal{W} = [0, 1]$ and $\mathbb{P}_W$ is the uniform probability measure on $[0, 1]$.

Equation (8.9) defines a multi-valued mapping from $[0, 1]$ to $2^{\mathcal{X}\times\Theta}$ that maps each W to the pairs (X, θ) compatible with W,

$$\Gamma : W \to \Gamma(W) = \{(x, \theta) \in \mathcal{X} \times \Theta | x = \varphi(\theta, W)\}. \tag{8.13}$$

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 6.1.2).

Assume that θ 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 (8.13) becomes

$$\Gamma_\theta : W \to \Gamma(W) = \{x \in \mathcal{X} | x = \varphi(\theta, W)\} = \{\varphi(\theta, W)\}. \tag{8.14}$$

By construction, for any measurable subset A of $\mathcal{X}$,

$$Bel^{\mathcal{X}}(A) = \mathbb{P}_W(\varphi(\theta, W) \in A) = \mathbb{P}_\theta(X \in A), \tag{8.15}$$

i.e., $Bel^{\mathcal{X}}$ is the probability distribution of X given θ.

Symmetrically, assume that we observe $X = x$. Our belief in θ is obtained by conditioning $Bel^{\mathcal{X}\times\Theta}$ on $\{x\} \times \Theta$ and marginalizing on Θ. The corresponding multi-valued mapping is

$$\Gamma_x : W \to \Gamma(W) = \{\theta \in \Theta | x = \varphi(\theta, W)\}. \tag{8.16}$$

The source $([0,1], \mathcal{B}([0,1]), \mathbb{P}_W, \Gamma_x)$ defines a belief function Bel_x^Θ on Θ, which represents the statistical evidence.

This method has the following important property. Let $X_1, \ldots, X_n$ be an iid sample from $f_\theta(x)$. Each X_i can be written as $X_i = \varphi(\theta, W_i)$, where $W_1, \ldots, W_n$ are iid from the standard uniform distribution. The belief function on Θ after observing $\mathbf{x} = (x_1, \ldots, x_n)$ is obtained by combining each $Bel_{x_i}^\Theta$ by Dempster's rule,

$$Bel_{\mathbf{x}}^\Theta = Bel_{x_1}^\Theta \oplus Bel_{x_2}^\Theta \oplus \ldots \oplus Bel_{x_n}^\Theta. \tag{8.17}$$

8.2.2 Application to a Bernoulli sample

Let us assume that we have an urn with an unknown proportion θ 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 $X_1, \ldots, X_n$, 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 [14]) is described in this section.

The φ-equation (8.9) in this case is

$$X_i = \varphi(\theta, W_i) = \begin{cases} 1 & \text{if } W_i \le \theta \\ 0 & \text{otherwise,} \end{cases} \tag{8.18}$$

where W_i has a standard uniform distribution. The multi-valued mapping Γ from (8.14) is

$$\Gamma(W) = (\{0\} \times [0, W)) \cup (\{1\} \times [W, 1]). \tag{8.19}$$

The form of the corresponding focal sets is illustrated in Figure 8.2.

After conditioning on x, we get

$$\Gamma_x(W) = \begin{cases} [0, W) & \text{if } x = 0, \\ [W, 1] & \text{if } x = 1, \end{cases} \tag{8.20}$$

which defines a random interval (see Section 6.2.2). The corresponding commonality function can be computed as follows. If $x = 0$, for any $0 \le u \le v \le 1$,

$$Q_x([u, v]) = \mathbb{P}_W([0, W) \supseteq [u, v]) = \mathbb{P}_W(W > v) = 1 - v. \tag{8.21}$$

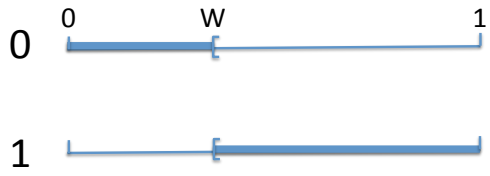


Figure 8.2: Focal sets $(\{0\} \times [0, W)) \cup (\{1\} \times [W, 1])$ of the belief function $Bel^{\mathcal{X}\times\Theta}$ in the case of a Bernoulli sample.

If $x = 1$,

$$Q_x([u, v]) = \mathbb{P}_W([w, 1] \supseteq [u, v]) = \mathbb{P}_W(W \leq u) = u. \tag{8.22}$$

After observing n realizations $x_1, \ldots, x_n$, the commonality function becomes, from (8.17) and (6.35),

$$Q([u, v]) \propto \prod_{i=1}^{n} Q_{x_i}([u, v]) = u^N (1 - v)^{n-N}, \tag{8.23}$$

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 (6.31), by differentiating $Q([u, v])$. If $0 < N < n$,

$$f(u, v) = -\frac{\partial^2 Q([u, v])}{\partial u \partial v} = c\, u^{N-1} (1 - v)^{n-N-1}, \tag{8.24}$$

where the proportionality constant can be shown to be

$$c = N(n - N) C_n^N. \tag{8.25}$$

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

$$Q([u, v]) = (1 - v)^n. \tag{8.26}$$

(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

$$f(v) = -\frac{\partial^2 Q([0, v])}{\partial v} = (1 - v)^{n-1}. \tag{8.27}$$

If $N = n$, the focal sets are of the form $[u, 1]$ and

$$Q([u, v]) = u^n. \tag{8.28}$$

The lower bound V is constantly equal to 1, while the lower bound has density

$$f(u) = -\frac{\partial^2 Q([u,1])}{\partial u} = nu^{n-1}. \tag{8.29}$$

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 θ, and $n-N$ are strictly greater than θ. 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 $W_1, \ldots, W_n$. 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 θ are thus easily computed from (6.24) and (6.25) as

$$Bel(\theta \leq t) = F_{B(N,n-N+1)}(t), \tag{8.30a}$$

$$Pl(\theta \leq t) = F_{B(N+1,n-N)}(t), \tag{8.30b}$$

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 θ are obtained from (**??**) and (**??**) as

$$\mathbb{E}_*(\theta) = \mathbb{E}(U) = \frac{N}{n+1}, \tag{8.31a}$$

$$\mathbb{E}^*(\theta) = \mathbb{E}(V) = \frac{N+1}{n+1}. \tag{8.31b}$$

8.3 Likelihood-based method

Although based on simple principles, Dempster's method of inference quickly leads to intricate derivations, as shown in Section 8.2.2, where it was applied to one the of 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 [58], was later studied by Wasserman [76] and Aickin [1], among others. It was recently justified axiomatically by Denœux [22].

8.3.1 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 θ by a belief function Bel_x^Θ. Which requirements should be imposed on Bel_x^Θ? The following three requirements seem relevant [22]:

1. Likelihood principle: Bel_x^Θ should only depend on the likelihood function. As noted in Section 8.1.3, the likelihood principle has a solid foundation, as it can be derived from the two generally accepted principles of exhaustivity and conditionality [10].

2. Compatibility with Bayesian inference: if a Bayesian prior $\pi(\theta)$ is available, combining it with Bel_x^Θ using Dempster's rule should yield the Bayesian posterior.

3. Least Commitment Principle (see Chapter 4): Bel_x^Θ 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^Θ should be proportional to the likelihood function:

$$pl_x(\theta) \propto L_x(\theta), \tag{8.32}$$

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 (8.32), is the consonant belief function whose contour function is the relative likelihood function,

$$pl_x(\theta) = \frac{L_x(\theta)}{\sup_{\theta' \in \Theta} L_x(\theta')}. \tag{8.33}$$

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 Θ induced by x. The corresponding plausibility function can be computed from pl_x as:

$$Pl_x^\Theta(A) = \sup_{\theta \in A} pl_x(\theta), \tag{8.34}$$

for all $A \subseteq \Theta$. The focal sets of Bel_x^Θ are the levels sets of $pl_x(\theta)$ defined as follows:

$$\Gamma_x(S) = \{\theta \in \Theta | pl_x(\theta) \geq S\}, \tag{8.35}$$

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^Θ (see Section 6.2.1). In particular, the following equalities hold:

$$Bel_x^\Theta(A) = \mathbb{P}_S(\{\Gamma_x(S) \subseteq A\}) \tag{8.36a}$$

$$Pl_x^\Theta(A) = \mathbb{P}_S(\{\Gamma_x(S) \cap A \neq \emptyset\}), \tag{8.36b}$$

for all $A \subseteq \Theta$ such that the above expressions are well-defined.

We can also remark that the maximum likelihood estimate (MLE) of θ can be interpreted as the value of θ with the highest plausibility, and likelihood regions as defined by Edwards [27], among others, are identical to plausibility regions.

8.3.2 Bernoulli example

As an example, let us consider the same model as in Section 8.2.2. The likelihood function after observing a realization $\mathbf{x} = (x_1, \ldots, x_n)$ of the iid random sample $X_1, \ldots, X_n$ is

$$L_{\mathbf{x}}(\theta) = \theta^N (1-\theta)^{n-N} \tag{8.37}$$

with $N = \sum_{i=1}^{n} x_i$. The likelihood-based belief function induced by x has the following contour function:

$$pl_{\mathbf{x}}(\theta) = \frac{\theta^x (1-\theta)^{n-N}}{\widehat{\theta}^N (1-\widehat{\theta})^{n-N}} = \left(\frac{\theta}{\widehat{\theta}}\right)^{n\widehat{\theta}} \left(\frac{1-\theta}{1-\widehat{\theta}}\right)^{n(1-\widehat{\theta})}, \tag{8.38}$$

for all $\theta \in \Theta = [0,1]$, where $\widehat{\theta} = N/n$ is the maximum likelihood estimate (MLE) of θ. Function $pl_{\mathbf{x}}(\theta)$ is plotted in Figure 8.3 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 $[U, V]$ [18]. The marginal cumulative probability distribution of U and V can be obtained as follows:

$$\begin{aligned}
F_U(u) &= \mathbb{P}(U \leq u) && \text{(8.39a)}\\
&= \mathbb{P}([U,V] \cap (-\infty, u] \neq \emptyset) && \text{(8.39b)}\\
&= Pl_{\mathbf{x}}^{\Theta}((-\infty, u]) && \text{(8.39c)}\\
&= \begin{cases} pl_{\mathbf{x}}(u) & \text{if } u \leq \widehat{\theta} \\ 1 & \text{otherwise,} \end{cases} && \text{(8.39d)}
\end{aligned}$$

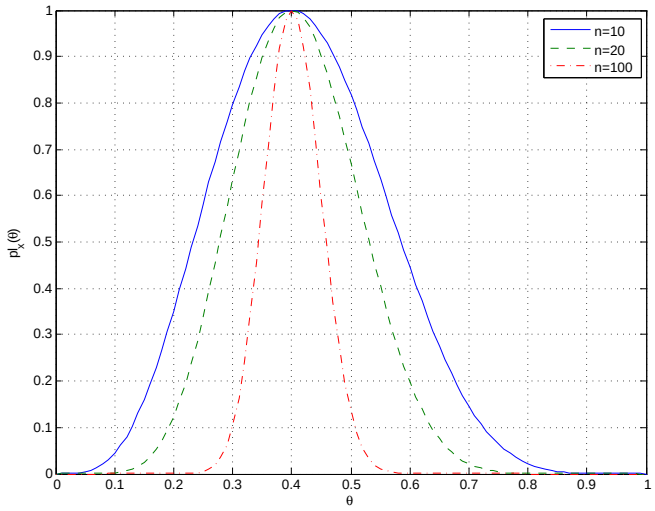


Figure 8.3: Contour functions (normalized likelihood functions) for the binomial distribution with $\widehat{\theta} = 0.4$ and $n \in \{10, 20, 100\}$.

and

$$
\begin{aligned}
F_V(v) &= \mathbb{P}(V \leq v) && \text{(8.40a)}\\
&= \mathbb{P}([U, V] \subseteq (-\infty, v]) && \text{(8.40b)}\\
&= Bel^{\Theta}_{\mathbf{x}}((-\infty, v]) && \text{(8.40c)}\\
&= 1 - Pl^{\Theta}_{\mathbf{x}}((v, +\infty)) && \text{(8.40d)}\\
&= \begin{cases} 0 & \text{if } v \leq \widehat{\theta} \\ 1 - pl_{\mathbf{x}}(v) & \text{otherwise.} \end{cases} && \text{(8.40e)}
\end{aligned}
$$

8.3.3 Properties

Viewing the relative likelihood function as the contour function of a consonant belief function or, equivalently, as a possibility distribution [82, 26] is, to a large extent, consistent with statistical practice. For instance, likelihood intervals [32, 70] 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 Θ_1 is

$$
pl_x(\theta_1) = \sup_{\theta_2 \in \Theta_2} pl_x(\theta_1, \theta_2), \tag{8.41}
$$

which is the relative profile likelihood function when θ_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

$$\Lambda(x) = \frac{\sup_{\theta \in H_0} L_x(\theta)}{\sup_{\theta' \in \Theta} L_x(\theta')} = \sup_{\theta \in H_0} pL_x(\theta) = Pl_x^{\Theta}(H_0). \tag{8.42}$$

However, the likelihood-based construction of belief functions is not compatible with Dempster's rule, i.e., Property (8.17) does not hold for this method. This does seem to be a weak point of this approach [61], although one might as well question Dempster's rule for combining independent statistical evidence, as was done by Aickin [1] and Walley [74], among others. Let E and E' be two independent random experiments with the same parameter space Θ, producing outcomes x and y according to frequency distributions $f_\theta(x)$ and $g_\theta(y)$. Let Bel_x^{Θ} and Bel_y^{Θ} denote the belief functions on Θ obtained after observing x and y, respectively. It is clear that $Bel_x^{\Theta} \oplus Bel_y^{\Theta}$ and Bel_{xy}^{Θ} are different, although they have the same contour function. However, Bel_{xy}^{Θ} can be obtained from Bel_x^{Θ} and Bel_y^{Θ} using the product rule of Possibility theory [26], which amounts to multiplying the contour functions (or possibility distributions) and renormalizing:

$$pl_{xy}(\theta) = \frac{pl_x(\theta) pl_y(\theta)}{\sup_{\theta' \in \Theta} pl_x(\theta') pl_y(\theta')}. \tag{8.43}$$

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 [23].

8.4 Prediction

In Section 8.2 and 8.3, we have described two solutions to the estimation problem, which consists in making statements about some parameter θ 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 θ 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 θ. In some cases, $X = (X_1, \ldots, X_n)$ and $Y = X_{n+1}$, where $X_1, \ldots, X_n, X_{n+1}$ is an iid sample. However, the model used here is more general. For instance, $X = (Z_0, Z_1, \ldots, Z_T)$ might be a time series and $Y = (Z_{T+1}, \ldots, Z_{T+h})$ 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 [37, 36].

8.4.1 General method

To make statements about Y, given some partial knowledge about θ, 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 φ-equation such as (8.9),

$$Y = \varphi'(\theta, W'), \tag{8.44}$$

where, as before, we will assume, without loss of generality, W' to have a standard uniform distribution. This equation defines a multi-valued mapping

$$\Gamma' : W \to \Gamma'(W) = \{(y, \theta) \in \mathcal{Y} \times \Theta | y = \varphi'(\theta, W')\}, \tag{8.45}$$

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^{Θ} and $Bel^{\mathcal{Y}\times\Theta}$, induced by multi-valued mapping $S \to \Gamma_x(S)$ and $W' \to \Gamma'(W')$. 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^{Θ} and $Bel^{\mathcal{Y}\times\Theta}$ using Dempster's rule and marginalizing on $\mathcal{Y}$,

$$Bel_x^{\mathcal{Y}} = \left(Bel_x^{\Theta} \oplus Bel^{\mathcal{Y}\times\Theta}\right)^{\downarrow\mathcal{Y}}. \tag{8.46}$$

The corresponding multi-valued mapping is

$$\Gamma_{\cap}(S, W') = \left[(\mathcal{Y} \times \Gamma_x(S)) \cap \Gamma'(W')\right]^{\downarrow\mathcal{Y}} \tag{8.47a}$$
$$= \varphi'(\Gamma_x(S), W'). \tag{8.47b}$$

We then have, for any measurable $A \subseteq \mathcal{Y}$,

$$Bel_x^{\mathcal{Y}}(A) = \mathbb{P}_{S,W'}\left(\varphi'(\Gamma_x(S), W') \subseteq A\right), \tag{8.48a}$$
$$Pl_x^{\mathcal{Y}}(A) = \mathbb{P}_{S,W'}\left(\varphi'(\Gamma_x(S), W') \cap A \neq \emptyset\right), \tag{8.48b}$$

where $\mathbb{P}_{S,W'}$ is the product measure $\mathbb{P}_S \otimes \mathbb{P}_{W'}$.

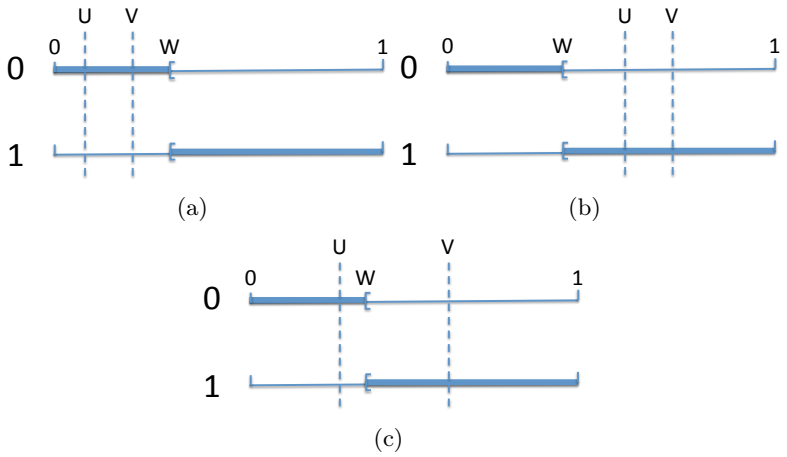


Figure 8.4: Three cases in the computation of the predictive belief function on Y in the Bernoulli example.

8.4.2 Bernoulli example

Let us return to the Bernoulli example considered in Sections 8.2.2 and 8.3.2 and let us address the problem of predicting a new value Y drawn independently from the Bernoulli $\mathcal{B}(\theta)$ distribution. We have

$$Y = \varphi'(\theta, W') = \begin{cases} 1 & \text{if } W' \leq \theta \\ 0 & \text{otherwise,} \end{cases} \tag{8.49}$$

where W' has a standard uniform distribution.

Each of the two estimation methods studied in Sections 8.2 and 8.3 provides a random interval $\Gamma_x(S) = [U(S), V(S)]$ on Θ. Therefore, (8.47) becomes

$$\big[(\{0,1\} \times [U,V]) \cap \Gamma(W')\big]^{\downarrow\{0,1\}} = \begin{cases} \{0\} & \text{if } [U,V] \subseteq [0,W), \\ \{1\} & \text{if } [U,V] \subseteq [W,1], \\ \{0,1\} & \text{otherwise.} \end{cases} \tag{8.50}$$

The three cases in (8.50) are illustrated in Figures (8.4(a)), (8.4(b)) and

(8.4(c)). It follows from (8.50) that the predictive mass function $m^{\mathcal{Y}}$ is

$$m^{\mathcal{Y}}(\{0\}) = P([U,V] \subseteq [0,W')) \tag{8.51a}$$
$$= \int P(V < W'|V = v)f(v)dv \tag{8.51b}$$
$$= \int (1-v)f(v)dv = 1 - \mathbb{E}(V), \tag{8.51c}$$

$$m^{\mathcal{Y}}(\{1\}) = P([U,V] \subseteq [W',1]) \tag{8.52a}$$
$$= \int P(U \geq W'|U = u)f(u)du \tag{8.52b}$$
$$= \int uf(v)dv = \mathbb{E}(U), \tag{8.52c}$$

and

$$m^{\mathcal{Y}}(\{0,1\}) = \mathbb{E}(V) - \mathbb{E}(U). \tag{8.53}$$

Equivalently, we have

$$Bel^{\mathcal{Y}}(\{1\}) = m^{\mathcal{Y}}(\{1\}) = \mathbb{E}(U) \tag{8.54}$$

and

$$Pl^{\mathcal{Y}}(\{1\}) = m^{\mathcal{Y}}(\{1\}) + m^{\mathcal{Y}}(\{0,1\}) = \mathbb{E}(V). \tag{8.55}$$

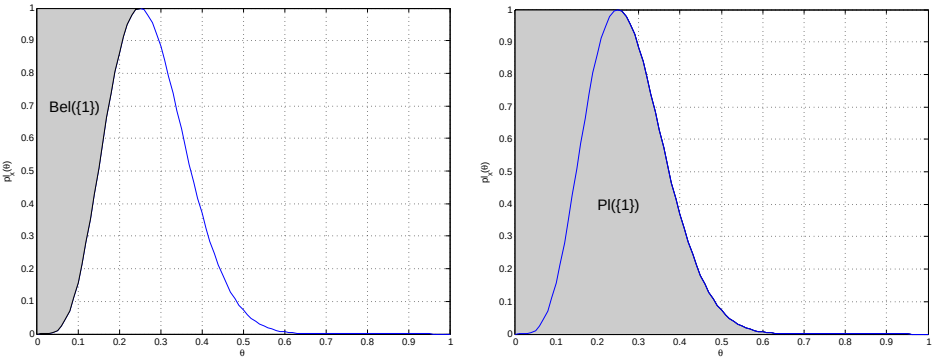
When $[U,V]$ is computed using Dempster's method, the expectations of U and V are given by (8.31). We then have

$$Bel^{\mathcal{Y}}(\{1\}) = \frac{N}{n+1}, \tag{8.56}$$
$$Pl^{\mathcal{Y}}(\{1\}) = \frac{N+1}{n+1}, \tag{8.57}$$

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 (**??**),

$$\mathbb{E}(U) = \int_0^{+\infty} (1 - F_U(u))du \tag{8.58a}$$
$$= \int_0^{\widehat{\theta}} (1 - pl_x(u))du \tag{8.58b}$$
$$= \widehat{\theta} - \int_0^{\widehat{\theta}} pl_x(u)du \tag{8.58c}$$



(a) (b)

Figure 8.5: Predictive belief and plausibility of success for a Bernoulli trial based on the contour function $pl_x(\theta)$ on the probability of success θ.

and

$$\mathbb{E}(U) = \int_0^{+\infty} (1 - F_V(v))du \tag{8.58d}$$

$$= \widehat{\theta} + \int_{\widehat{\theta}}^{1} pl_x(v)dv. \tag{8.58e}$$

These two quantities can be represented as the areas of regions delimited by the contour function, as shown in Figure 8.5. 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.

8.4.3 Monte Carlo approximation

In practice, the analytical expression of the predictive belief function induced by the multi-valued mapping (8.47) may be intractable. We can then resort to Monte Carlo approximation, by drawing independently n pairs $(S_1, W_1), \ldots, (S_n, W_n')$ 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 (8.48) by, respectively,

$$\widehat{Bel}_x^{\mathcal{Y}}(A) = \frac{1}{n}\#\{\varphi'(\Gamma_x(S_i), W_i') \subseteq A\}, \tag{8.59a}$$

$$\widehat{Pl}_x^{\mathcal{Y}}(A) = \frac{1}{n}\#\{\varphi'(\Gamma_x(S_i), W_i') \cap A \neq \emptyset\}. \tag{8.59b}$$

If the mapping $\theta \to \varphi'(\theta, w)$ 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:

$$a_i = \min \varphi'(\theta, W_i) \tag{8.60a}$$

and

$$b_i = \max \varphi'(\theta, W_i) \tag{8.60b}$$

under the constraint

$$L_x(\theta) \geq S_i. \tag{8.60c}$$

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,

$$\mathbb{E}_*(Y) \approx \frac{1}{n}\sum_{i=1}^{n} a_i, \quad \mathbb{E}^*(Y) \approx \frac{1}{n}\sum_{i=1}^{n} b_i. \tag{8.61}$$

Similarly, the lower and lower predictive quantiles of Y (see Section 6.2.2) can be approximated by the empirical quantiles of the a_i's and b_i's.

8.4.4 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^{Θ} by Dempster's rule. As mentioned in Section 8.3.1, the combined belief function $Bel_x^{\Theta} \oplus \pi$ is then, by construction, the posterior probability distribution $f_x(\theta)$ of θ given $X = x$ and we then have, for any measurable subset $A \subseteq \mathcal{Y}$:

$$\begin{aligned}
Bel_x^{\mathcal{Y}}(A) &= \mathbb{P}(\varphi'(\theta, W') \in A | x) && \text{(8.62a)}\\
&= \int_{\Theta} \mathbb{P}(\varphi'(\theta, W') \in A | \theta, x) f_x(\theta) d\theta && \text{(8.62b)}\\
&= \int_{\Theta} \left(\int_A g_{x,\theta}(y) dy \right) f_x(\theta) d\theta && \text{(8.62c)}\\
&= \int_A \left(\int_{\Theta} g_{x,\theta}(y) f_x(\theta) d\theta \right) dy && \text{(8.62d)}\\
&= \int_A g_x(y) dy, && \text{(8.62e)}
\end{aligned}$$

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 9

Classification and clustering

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] R. G. Almond. *Graphical belief models.* Chapman and Hall, London, 1995.

[4] 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.

[5] 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.

[6] J. A. Barnett. Computational methods for a mathematical theory of evidence. In *Proceedings IJCAI-81*, pages 868–875, Vancouver, 1981.

[7] 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.

[8] 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.

[9] J. M. Bernardo and A. F. M. Smith. *Bayesian theory.* John Wiley and Sons, New-York, 1994.

[10] A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

[11] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–12, 1946.

[12] B. de Finetti. La prévision : ses lois logiques, ses sources objectives. *Annales de l'Institut Henri Poincarré*, 7(1):1–68, 1937.

[13] M. A. B. Deakin. The wine/water paradox: background, provenance and proposed resolutions. *The Australian Mathematical Society Gazette*, 33(3):200–205, 2006.

[14] A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374, 1966.

[15] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[16] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 57:515–528, 1967.

[17] A. P. Dempster. A generalization of Bayesian inference (with discussion). *J. R. Statistical Society B*, 30:205–247, 1968.

[18] A. P. Dempster. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics*, 39(3):957–966, 1968.

[19] A. P. Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.

[20] 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.

[21] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

[22] T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.

[23] 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.

[24] 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.

[25] 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.

[26] D. Dubois and H. Prade. *Possibility Theory: An approach to computerized processing of uncertainty*. Plenum Press, New-York, 1988.

[27] A. W. F. Edwards. *Likelihood (expanded edition)*. The John Hopkins University Press, Baltimore, USA, 1992.

[28] D. Ellsberg. Risk, ambiguity and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

[29] R. A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[30] R. A. Fisher. *Statistical methods and scientific inference*. Oliver and Boyd, Edinburgh, 1956.

[31] D. Harmanec. Faithful approximations of belief functions. In K. B. Laskey and H. Prade, editors, *Uncertainty in Artificial Intelligence 15 (UAI99)*, Stockholm, 1999.

[32] D. J. Hudson. Interval estimation from the likelihood function. *J. R. Statistical Society B*, 33(2):256–262, 1973.

[33] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied Interval Analysis*. Springer-Verlag, London, 2001.

[34] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003.

[35] 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.

[36] 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.

[37] O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Forecasting using belief functions: an application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.

[38] R. Kennes. Computational aspects of the Möbius transform of graphs. *IEEE Trans. SMC*, 22:201–223, 1992.

[39] F. Klawonn and P. Smets. The dynamic of belief in the transferable belief model and specialization-generalization matrices. In D. D. et al., editor, *Proc. of the 8th conference on Uncertainty in Artificial Intelligence*, pages 130–137. Morgan Kaufmann, San Mateo, CA, 1992.

[40] G. J. Klir and M. J. Wierman. *Uncertainty-Based Information. Elements of Generalized Information Theory*. Springer-Verlag, New-York, 1999.

[41] J. Kohlas and P.-A. Monney. *A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory of Evidence*. Springer-Verlag, Berlin, 1995.

[42] 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.

[43] 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.

[44] R. D. Luce and H. Raiffa. *Games and Decisions: Introduction and Critical Survey*. Wiley, New York, 1957.

[45] J. M. Mikkelson. Dissolving the wine/water paradox. *British Journal for the Philosophy of Science*, 55(1):137–145, 2004.

[46] P.-A. Monney. *A Mathematical Theory of Arguments for Statistical Evidence*. Contributions to Statistics. Physica-Verlag, Heidelberg, 2003.

[47] R. E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New-Jersey, 1966.

[48] R. E. Moore, R. B. Kearfott, and M. J. Cloud. *Introduction to Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[49] 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.

[50] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

[51] H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC Press, Boca Raton, Florida, 2006.

[52] B. W. Pearson, III. *Introduction to Management Science*. Prentice Hall, Upper Saddle River, NJ, USA, 12th edition, 2014.

[53] K. R. Popper. The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:25–42, 1959.

[54] A. Ramer. Uniqueness of information measure in the theory of evidence. *Fuzzy Sets and Systems*, 24:183–196, 1987.

[55] 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.

[56] L. J. Savage. *The foundations of statistics*. Wiley, New York, 1954.

[57] G. Shafer. *Allocations of probability: A theory of partial belief*. PhD thesis, Princeton University, 1973.

[58] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[59] G. Shafer. Allocations of probability. *Annals of Probability*, 7(5):827–839, 1979.

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

[61] G. Shafer. Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B*, 44:322–352, 1982.

[62] G. Shafer, P. P. Shenoy, and K. Mellouli. Propagating belief functions in qualitative Markov trees. *International Journal of Approximate Reasoning*, 1:349–400, 1987.

[63] P. P. Shenoy. A valuation-based language for expert systems. *International Journal of Approximate Reasoning*, 3:383–411, 1989.

[64] P. P. Shenoy. Binary joint trees for computing marginals in the Shenoy-Shafer architecture. *International Journal of Approximate Reasoning*, 17:239–263, 1997.

[65] P. Smets. Information content of an evidence. *International Journal of Man-Machine Studies*, 19:33–43, 1983.

[66] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[67] 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.

[68] P. Smets. The application of the matrix calculus to belief functions. *International Journal of Approximate Reasoning*, 31(1–2):1–30, 2002.

[69] P. Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8(4):387–412, 2007.

[70] D. A. Sprott. *Statistical Inference in Science*. Springer-Verlag, Berlin, 2000.

[71] K. Szaniawski. Some remarks concerning the criterion of rational decision making. *Studia Logica*, 9(1):221–239, 1960.

[72] B. Tessem. Approximations for efficient computation in the theory of evidence. *Artificial Intelligence*, 61:315–329, 1993.

[73] J. von Neumann and O. Morgenstern. *Theory Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

[74] P. Walley. Belief function representations of statistical evidence. *The Annals of Statistics*, 15(4):1439–1465, 1987.

[75] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[76] L. A. Wasserman. Belief functions and statistical evidence. *The Canadian Journal of Statistics*, 18(3):183–196, 1990.

[77] P. M. Williams. Indeterminate probabilities. In M. Przelecki, K. Szaniawski, R. Wójcicki, and G. Malinowski, editors, *Formal Methods in the Methodology of Empirical Sciences*, volume 103 of *Synthese Library*, pages 229–246. Springer Netherlands, 1976.

[78] 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.

[79] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denœux. Multimodal information for urban scene understanding. *Machine Vision and Applications (to appear)*, 2015.

[80] R. R. Yager. The entailment principle for Dempster-Shafer granules. *Int. J. of Intelligent Systems*, 1:247–262, 1986.

[81] L. A. Zadeh. Fuzzy sets. *Inform. Control*, 8:338–353, 1965.

[82] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

Index