1

Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction*

Liping Liu and Ronald R. Yager

Abstract. In this chapter, we review the basic concepts of the theory of belief functions and sketch a brief history of its conceptual development. We then provide an overview of the classic works and examine how they established a body of knowledge on belief functions, transformed the theory into a computational tool for evidential reasoning in artificial intelligence, opened up new avenues for applications, and became authoritative resources for anyone who is interested in gaining further insight into and understanding of belief functions.

1 Introduction

The Dempster-Shafer theory of belief functions was due to the seminal work of Glenn Shafer and its conceptual forerunner—lower and upper probabilities by Arthur P. Dempster. This year marks respectively the 30th and 40th anniversaries of these two important publications. In the last 30 years, belief functions have penetrated into many scientific areas, technological projects, and educational enterprises. By bridging fuzzy logic and probabilistic reasoning, the theory of belief functions has become a primary tool for knowledge representation and uncertain reasoning in expert systems. Thanks to the availability of powerful computers and user-friendly software, belief functions have been widely applied to business, engineering, and medical problems. The applications include auditing, process engineering, quality control, decision support, electronic commerce, financial asset evaluation, information fusion, information retrieval, knowledge management, medical diagnosis, mobile services, natural resource detection, network security, object classification, risk management, software engineering, target tracking, etc.

To celebrate the anniversaries, to showcase the achievements, and to assess the current state of knowledge, the editors bring together a volume of 29 classic papers on the theory of belief functions and its applications. The collection was

* The authors would like to thank Glenn Shafer for his invaluable comments on earlier versions of this chapter.

created from a pool of over 100 nominated contributions, which are regarded as classics with a high prospect to continue to influence the future development of the field.

In this chapter, we introduce the Dempster-Shafer theory and present its basic concepts and major results. The goal is to summarize Glenn Shafer's classic book [34] in a concise, comprehensive, and accessible manner so that the reader will gain sufficient conceptual background to pursue further readings. Then we sketch a brief history of the earlier conceptual development, from Ronald A. Fisher's fiducial arguments to Arthur P. Dempster's generalized Bayesian inference, and from Jakob Bernoulli's notion of pure evidence to Glenn Shafer's mathematical theory of evidence. The goal is to expose the origin of the concepts so that the reader will gain a broad perspective for understanding further development. Then we provide an overview of the classic works and point out their unique contributions in terms of how they established a body of knowledge on belief functions, transformed the theory into a computational tool for evidential reasoning in artificial intelligence, opened up new avenues for applications in business, engineering, and medicine, and became authoritative resources for anyone who is interested in gaining further insight into and understanding of the theory. Finally, we briefly discuss famous critiques by Lotfi A. Zadeh and Judea Pearl and point out a few open problems that need to be solved in future research.

2 Basic Concepts

The concept of belief functions may be formalized in various ways. In this section, we adopt the approach by Glenn Shafer in his seminal work—*A Mathematical Theory of Evidence* [34]—for exposition since its terminologies and notations are the standard in the literature.

Given a question of interest, let Θ be a finite set of possible answers to the question, called a *frame of discernment*, and 2^Θ be the set of all subsets of Θ:

$$2^\Theta = \{A \mid A \subseteq \Theta\}.$$

The subset A includes as special cases the empty set ϕ and the full set Θ. It represents a statement or proposition that the truth lies in A. A real function over the subsets $Bel : 2^\Theta \rightarrow [0, 1]$ is called a *belief function* if and only if it satisfies the following three axioms:

Axiom 1 $Bel(\phi) = 0$.

Axiom 2 $Bel(\Theta) = 1$.

Axiom 3 *For any whole number* n *and subsets* $A_1, A_2, .., A_n \subset \Theta$,

$$Bel(\bigcup_{i=1}^{n} A_i) \geq \sum_{\substack{I \subset \{1,2,...,n\} \\ I \neq \phi}} (-1)^{|I|+1} Bel(\bigcap_{i \in I} A_i).$$

In the case where $n = 2$ and $A_1 \cap A_2 = \phi$, Axiom 3 reduces to $Bel(A_1 \cup A_2) \geq Bel(A_1) + Bel(A_2)$. The student of probability theory may immediately recognize that these axioms are similar to those for a probability function with the inequality of Axiom 3 substituting for equality. When equality holds, $Bel(A_1 \cup A_2) = Bel(A_1) + Bel(A_2)$ if $A_1 \cap A_2 = \phi$. Thus, a probability function is additive whereas a belief function is generally not. The generalized axioms, however, indicate that belief functions include probability functions as special cases and may be equally or better used to express degrees of belief.

Additive probabilities are common sense. Are there any non-additive beliefs to justify an extension? The answer is affirmative. A modern example is from Bayesian statistics on how to represent ignorance, where the truth is in Θ but there is no information, probabilistic or logical, to justify the whereabouts of the truth. Thus, $Bel(\Theta) = 1$ but $Bel(A) = 0$ for any proper subset A of Θ. Clearly, this function fails to be additive. An ancient example was due to Jakob Bernoulli in his book *Ars Conjectandi*. Suppose a man was stabbed with a sword in a milling crowd and Gracchus was interrogated and turned pale. Since the sign of pallor betokens a finite number of reasons: melancholy, fear, cold, anger, amorous passion, etc., it proves Gracchus guilty if it arises from a guilty conscience, but does not prove his innocence if it arises from other reasons. Thus, $Bel(\{guilty\}) < 1$ and $Bel(\{innocent\}) = 0$, but $Bel(\Theta) = 1$. Again, this belief function is non additive.

The key to the concept of belief functions is limited division of belief. Whereas probability functions assume belief is apportioned to the points in the frame Θ, belief functions allow basic probability numbers (or mass numbers), to be assigned to whole sets of points in Θ without further subdivision. The basic idea is that a whole belief is divided into one or more basic probability numbers $m(A)$ and allocated to one or more subsets A, called *focal elements*, such that:

$$\sum\{m(A) \mid A \subseteq \Theta\} = 1. \tag{1}$$

The basic probability number $m(A)$ allocated to a focal element A is not further divided into smaller chunks allocated to proper subsets of A.

This suggests an alternative approach to the definition of a belief function. Given basic probability numbers $m(A)$, the belief $Bel(A)$ is defined by:

$$Bel(A) = \sum\{m(B) \mid B \subseteq A\}. \tag{2}$$

Logically, a portion of belief committed to one proposition is committed to any other proposition it implies. Thus, the total belief committed to a subset A is the sum of those that are committed to proper subsets of A and those to A itself.

Adding the boundary condition $m(\phi) = 0$ to (1), Shafer showed that the two definitions are equivalent, i.e., a function satisfies the three axioms if and only if it can be represented as the sum of basic probability numbers over

focal elements. In fact, given a belief function, one may construct such a basic probability number for each $A \subseteq \Theta$ using a Möbius transformation:

$$m(A) = \sum\{(-1)^{|A-B|} Bel(B) \mid B \subseteq A\}, \tag{3}$$

where $|A - B|$ is the cardinality of $A - B$, or a recursive deduction:

$$m(\phi) = 0, m(A) = Bel(A) - \sum\{m(B) \mid B \subset A\}.$$

Despite the equivalence, however, one should note that the axiomatic definition allows the establishment of the theory of belief functions with no reference to probabilities.

Due to the limited divisibility, belief not committed to $\overline{A}$, the negation of A, is not automatically committed to A. But it does make A more credible or plausible. Thus, it is intuitive to define a plausibility function $Pl(A)$ as the sum of beliefs not committed to $\overline{A}$:

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

Through (2), it is easy to see the interplay between basic probability numbers and plausibility numbers as follows:

$$Pl(A) = \sum\{m(B) \mid A \cap B \neq \phi\}. \tag{5}$$

For any proposition, its plausibility is no less than its committed belief, i.e., $Bel(A) \leq Pl(A)$. Thus, in his earlier works [8, 9, 10], Dempster called these functions respectively lower and upper probabilities. The terminology had caused some confusion and was abandoned by Shafer.

Belief functions are meant to be a representation of subjective beliefs. Unlike other alternative formalisms, however, belief functions represent the beliefs grounded on or supported by evidence. In fact, the idea of limited divisibility makes intuitive sense if one interprets a basic probability number $m(A)$ as a measure of evidential support to A. Given two subsets A and B in a frame of discernment, if B is a proper subset of A, then B represents a stronger proposition than A and requires stronger evidence to support it. Therefore, the evidence that supports A does not automatically support B and the belief $m(A)$ committed to A does not necessitate the commitment of a smaller number $m(B)$ to B.

Given a distinct piece of evidence, its support may be encoded as a list of mass numbers assigned to the corresponding focal elements. It may also be summarized as a belief or plausibility function over the frame of discernment.

When there exist multiple items of evidence, of course, it is necessary to combine them together. *Dempster's rule of combination* serves this purpose. In his original framework of a multivalued mapping that carries a probability measure into a system of upper and lower probabilities (see below), Dempster derived this rule of combining upper and lower probabilities based on

the assumption that two probability measures were independent [9]. In the axiomatic framework, Shafer adopted the rule as a definition for combining distinct or independent bodies of evidence. Let m_1 and m_2 be the mass functions for two independent bodies of evidence. The combination via Dempster's rule follows a simple three-step process: intersection of focal elements, multiplication of corresponding basic probability numbers $m_1 m_2$, and normalization in accordance with (1). Each intersection, if it is not empty, becomes a new focal element of the combined belief function. The corresponding product of basic probability numbers contributes to the support to the new focal element. An empty intersection indicates a disagreement or conflict and is excluded from further consideration. Its corresponding product of basic probability numbers is subtracted from the whole belief mass for normalization. Mathematically, the new mass function over new focal elements is defined as follows:

$$m(A) = \frac{\sum\{m_1(B)m_2(C) \mid B \cap C = A\}}{\sum\{m_1(B)m_2(C) \mid B \cap C \neq \phi\}}. \tag{6}$$

Since empty intersection indicates a conflict, $\sum\{m_1(B)m_2(C) \mid B \cap C = \phi\}$ measures the total amount of conflict. Formally, we call the logarithm of the renormalization constant *the weight of conflict*:

$$W = \log(\frac{1}{\sum\{m_1(B)m_2(C) \mid B \cap C \neq \phi\}}). \tag{7}$$

Of course, two belief functions are *combinable* if and only if their weight of conflict is finite.

Example 1. Suppose, among three suspects, Tony (T), Smith (S), and Dick (D), we want to find out who committed a bank burglary. In the investigation, we questioned Mrs. Johnson, a witness who was living close to the bank. She said that she saw a big person near the bank around the time when the crime was committed. Assume Mrs. Johnson's testimony was 60% reliable based on her eyesight. If her testimony was reliable, the evidence pointed to Tony or Dick since they had big bodies. Thus, $m_1(\{T, D\}) = 0.6$. However, if she was not reliable, the testimony carried no information, i.e., $m_1(\{T, S, D\}) = 0.4$. Although the criminal wore a mask, a video camera recorded a fuzzy picture of the person's eyes, which were 4 times more likely to be black than to be gray. The second item of evidence suggested $m_2(\{S\}) = 0.8$ and $m_2(\{T, D\}) = 0.2$ since Smith had black eyes. To combine the two pieces of evidence, we can use a tabular form as in Table (1). For each cell, take the corresponding focal elements from each item of evidence, intersect them and multiply their corresponding basic probabilities. The weight of conflict between the two items of evidence is $\log(\frac{1}{1-0.48}) = 0.28$. In the combined evidence, there are two focal elements: {S} and {T, D}. The combined mass function is calculated as follows:

$$m(\{S\}) = \frac{0.32}{1 - 0.48} = 0.615,$$
$$m(\{T, D\}) = \frac{0.12 + 0.08}{1 - 0.48} = 0.385.$$

Table 1. An illustration of combination

	$m_2(\{S\}) = 0.8$	$m_2(\{T, D\}) = 0.2$
$m_1(\{T, D\}) = 0.6$	$\phi \to 0.48$	$\{T, D\} \to 0.12$
$m_1(\{T, S, D\}) = 0.4$	$\{S\} \to 0.32$	$\{T, D\} \to 0.08$

Thus Smith appeared to be more suspicious according to the combined evidence.

The combined belief and plausibility function may be symbolically expressed as $Bel_1 \oplus Bel_2$ and $Pl_1 \oplus Pl_2$, respectively. Unfortunately, there is no simple analytical expression for the orthogonal sum $\oplus$. To put the combination rule into multiplicative form as in the case for probability functions, Dempster introduced another function $Q(A)$, which Shafer called the *commonality function*, as follows:

$$Q(A) = \sum\{m(B) \mid B \supseteq A\}. \tag{8}$$

Let Q_1 and Q_2 be respectively the commonality functions for two independent items of evidence. Then the commonality function for the combined evidence is as follows:

$$Q(A) = \frac{Q_1(A)Q_2(A)}{\sum\{(-1)^{|A|+1}Q_1(A)Q_2(A) \mid A \neq \phi\}}. \tag{9}$$

Here the denominator is identical to that in (6).

Unlike belief and plausibility functions, a commonality function is not intuitive but Shafer interpreted $Q(A)$ as the total belief that is free to move to every element of A. According to (8) and (3), it is clear that the definition of a commonality function is opposite to that of a belief function in the sense that a belief for A sums all basic probability numbers committed to A and its proper subsets whereas a commonality number sums those that are committed to A and its proper supersets. Consequently, commonality functions are decreasing while belief (and plausibility) functions are increasing: for any two propositions A and B, if $A \supset B$, then $Q(A) \leq Q(B)$ but $Bel(A) \geq Bel(B)$ and $Pl(A) \geq Pl(B)$.

The four representations of evidence, namely, belief functions $Bel(A)$, mass functions $m(A)$, plausibility functions $Pl(A)$, and commonality functions $Q(A)$, are interrelated. Some of the relationships are shown below: for any non-empty set A,

$$\begin{aligned}
Bel(A) &= \sum\{(-1)^{|B|}Q(B) \mid B \subseteq \overline{A}\},\\
Q(A) &= \sum\{(-1)^{|B|}Bel(\overline{B}) \mid B \subseteq A\},\\
Pl(A) &= \sum\{(-1)^{|B|+1}Q(B) \mid \phi \neq B \subseteq A\},\\
Q(A) &= \sum\{(-1)^{|B|+1}Pl(B) \mid B \subseteq A\}.
\end{aligned}$$

From any representation one can obtain another one through a series of additions and/or Möbius transformations. In this sense, all the representations are equivalent. Thus, one may start with any one model to encode evidence and end up with other representations for decision making or probable reasoning. The choice is purely based on convenience. Mass functions are often a more natural and superior device for encoding evidence, whereas belief and plausibility functions are a more intuitive summary of the impact of the evidence on propositions. After all, evidence often arises in the form of knowledge in a related domain that provides insights on or connections to propositions in the domain of interest. If the knowledge is probabilistic, it can then be carried over to the propositions of interest as basic probability numbers. For example [19], suppose I find a scrap of newspaper predicting a blizzard, which I regard as infallible. Also, suppose I am 75% certain that the newspaper is today's. Here the knowledge about the newspaper maps to tomorrow's weather as follows: if the newspaper is today's, then a blizzard is sure to come; if the newspaper is not today's, however, it provides no information on tomorrow's weather. Thus, we transfer 75% as a basic probability number to the focal element {blizzard}, i.e., $m(\{\text{blizzard}\}) = 0.75$, and 25% to Θ, i.e., $m(\Theta) = 0.25$. Of course, there are occasions when belief or plausibility functions become more convenient. For example, Srivastava and Shafer [40] interpret audit risks as the plausibility that a financial statement is not fairly stated or an audit objective is not met. Thus it is more convenient to use plausibility functions to encode audit evidence.

To illustrate the equivalence of the four representations, Table 2 shows the respective representations of three special cases of belief functions, namely vacuous belief functions, Bayesian belief functions, and simple support functions. A *vacuous belief function* represents full ignorance, i.e., evidence does not provide any support to or information on any specific proposition, i.e., any proper subset of a frame of discernment. Thus, Θ is the only focal element. *A Bayesian belief function* represents probabilistic knowledge that assigns a probability to each element of Θ. In other words, all focal elements are singletons. *A simple support function* represents a piece of homogeneous evidence that provides support to one and only one proposition that is a proper subset of Θ. In other words, there are two focal elements: S and Θ with $S \subset \Theta$.

Despite their simplicity, the three special cases play important roles in the theory of belief functions in the sense that: 1) they are the building blocks for more complex belief functions; and 2) they justify the superiority of belief functions to probability theory. Vacuous belief functions provide a simple solution to the problem of representing ignorance. Note that Bayesian statistics would represent full ignorance as a uniform distribution, which essentially mixes lack of belief with disbelief. For example, what is my belief that a coin will land a head? It is 50% if and only if I know the coin is fair. If I am ignorant, the most I can say is $Bel(\Theta) = 1$. However, Bayesian statistics will assign 50% as a prior probability regardless.

Table 2. Three special cases

	Mass function	Belief function
Vacuous belief functions	$m(\Theta) = 1$	$Bel(A) = \begin{cases} 0 & \forall A \subset \Theta \\ 1 & A = \Theta \end{cases}$
Bayesian belief functions	$\|A\| = 1$ for each focal element A	$Bel(A)$ is additive
Simple support functions	$m(A) = \begin{cases} s & A = S \\ 1-s & A = \Theta \\ 0 & \text{else} \end{cases}$	$Bel(A) = \begin{cases} s & A \supseteq S \\ 1 & A = \Theta \\ 0 & \text{else} \end{cases}$
	Plausibility function	**Commonality function**
Vacuous belief functions	$Pl(A) = 1\ \forall A \neq \phi$	$Q(A) = 1\ \forall A$
Bayesian belief functions	$Pl(A)$ is additive	$Q(A) = 0$ if $\|A\| > 1$
Simple support functions	$Pl(A) = \begin{cases} 1 & A \cap S \neq \phi \\ 1-s & A \cap S = \phi \end{cases}$	$Q(A) = \begin{cases} 1 & A \subseteq S \\ 1-s & \text{else} \end{cases}$

Bayesian belief functions are regular probabilities. They are the only case where beliefs and plausibilities are identical, i.e., $Bel(A) = Pl(A)$ for any $A \subseteq \Theta$, and additive as well, i.e., $Bel(A_1 \cup A_2) = Bel(A_1) + Bel(A_2)$ if $A_1 \cap A_2 = \phi$. Thus, belief functions include probability functions as a special case. It is also the only case that we have zero commonality number for any subset of cardinality 2 or larger.

The concept of simple support functions is the most important extension to Dempster's work on generalized Bayesian inference. It acts as the basis for defining the *weight of evidence*, by which Bernoulli meant probative force for a probability judgment. For a simple support function with $m(S) = s$ and $m(\Theta) = 1 - s$, the weight of evidence w is a nonnegative number in $[0, \infty)$ that maps to the support s in such a way that the sum of two weights maps to the combined support of the two items of evidence via Dempster's rule. This along with the following boundary condition:

$$s = \begin{cases} 0 & w = 0 \\ 1 & w \to \infty \end{cases}$$

leads to an analytical expression of the weight of evidence:

$$w = -\log(1 - s).$$

Since a simple support function uniquely determines a weight of evidence, it is tempting to decompose a general belief function into one or more simple support functions and then derive the weight of evidence underlying it. Toward this goal, Shafer defined the concept of *a separable support function*

to be the orthogonal sum of one or more simple support functions. Unlike a simple support function, a separable support function may support multiple propositions that are proper subsets of Θ. Unlike a general belief function, it is distinct in that, for any two focal element A and B, if $A \cap B \neq \phi$, then $A \cap B$ is also a focal element.

As an example of special importance, *consonant support functions* are separable support functions. A belief function is called consonant if its focal elements are nested, i.e., for any two focal elements A and B, either $A \subset B$ or $B \subset A$. Thus, all focal elements may be arranged in an order of increasing precision, pointing in a single direction. A consonant support function Bel has the following distinct features:

$$\begin{aligned} Bel(A \cap B) &= \min(Bel(A), Bel(B)) \ \forall A, B \subseteq \Theta, \\ Pl(A \cup B) &= \max(Pl(A), Pl(B)) \ \forall A, B \subseteq \Theta, \\ Q(A) &= \min\{Q(\theta) \mid \theta \in A\} \ \forall A \neq \phi. \end{aligned}$$

Those familiar with fuzzy logic may recognize that the *possibility* and *necessity functions* introduced by Zadeh [42] are the same as consonant plausibility and support functions. A function f is a consonant support function if and only if it satisfies: $f(\phi) = 0$, $f(\Theta) = 1$, and $f(A \cap B) = \min(f(A), f(B))$ for any $A, B \subseteq \Theta$. These are the axioms used for developing the theory of possibility.

There is no unique way to decompose a separable support function into simple support functions. For example, one simple support function may be further represented as the orthogonal sum of two or more simple support functions that support the same proposition. If no component has infinite weight of evidence, however, this non-uniqueness does not cause any trouble because the total weight of evidence focused on each subset will be the same no matter which decomposition is used. Let S_i be the proposition supported by the ith component and w_i be the corresponding weight of evidence. If w_i is finite for all i, then the *total weight of evidence* focused on any non-empty proper subset A of Θ is

$$w(A) = \sum\{w_i \mid S_i = A\},$$

with $w(\phi) = 0$ and $w(\Theta) = \infty$.

Through a weight function $w(A)$, one may define two related concepts: the impingement function $v(A)$ and the weight of internal conflict W. The impingement function $v(A)$ is defined as the sum of the weights of evidence focused on the propositions not containing A:

$$v(A) = \sum\{w(B) \mid A \cap \overline{B} \neq \phi\}. \tag{10}$$

Each weight $w(B)$ impugns all propositions not in its focus B. Thus $v(A)$ is the total weight of evidence not favoring A. Given an impingement function, one may recover the weight function using a Möbius transformation, i.e., for each non-empty proper subset A of Θ,

$$w(A) = \sum\{(-1)^{|B-A|} v(B) \mid A \subseteq B\}.$$

The internal conflict of a separable support function refers to the conflict among the simple support functions that make up the separable support function. Its weight can be defined as in (7) with a straightforward extension to multiple belief functions. Since decomposition may not be unique, the weight of conflict in general varies from decomposition to decomposition. The *weight of internal conflict* is actually defined as the minimum of the weights of conflict for all possible decompositions. The weight of internal conflict can be expressed in terms of the impingement function $v(A)$ or the commonality function $Q(A)$:

$$W = -\log(\sum\{(-1)^{|A|+1} \exp(-v(A)) \mid A \neq \phi\},$$
$$W = -\sum\{(-1)^{|A|} \log Q(A) \mid A \subseteq \Theta\}.$$

The above equations give another way to express the commonality function $Q(A)$ for a separable support function as follows:

$$\log Q(A) = W - v(A). \tag{11}$$

The total weight of evidence determines the impingement function, which in turn determines the weight of internal conflict. Thus, it determines a commonality function, from which one can recover a mass function, a belief function, and a plausibility function. Therefore, for separable support functions, the total weight of evidence provides a sufficient assessment of evidence.

Equation (11) shows an intuitive association of smaller commonality numbers with greater degrees of impingement. Formally, suppose v_1 and v_2 are two impingement functions and Q_1 and Q_2 are the corresponding commonality functions. If $Q_1(A) \leq Q_2(A)$ for all $A \subseteq \Theta$, then $v_1(A) \geq v_2(A)$ for all $A \subseteq \Theta$. This association can easily be derived from the following still unproven *weight-of-conflict conjecture*: if Q_1 and Q_2 are the commonality functions for two separable support functions and W_1 and W_2 are their corresponding weights of internal conflicts, then

$$Q_1(A) \leq Q_2(A)\ \forall A \subseteq \Theta \implies W_1 \geq W_2. \tag{12}$$

In the axiomatic approach, any function that satisfies Axioms 1–3 is a belief function. A whole body of mathematical theory of belief functions could have been built based on these axioms. However, Shafer was interested in building a theory of evidence as a science of probable reasoning. Thus, his central theme was to investigate which subclasses of belief functions could be useful for the representation of evidence. As we have seen, both simple and separable support functions were proposed for such a purpose; they are or can be decomposed into components each of which precisely and homogeneously supports a given proposition.

Toward the same goal, the concept of support functions was proposed. A *support function* is a belief function that can be derived from the marginalization of a separable support function to a coarser framer of discernment. Unlike a *sample space* in probability theory, a frame of discernment is epistemic in nature and is constructed for probable reasoning. It can be refined or coarsened as needed. For example, suppose we are interested in whether tomorrow's weather will be raining (r), snowing (s), or normal (n). The frame of discernment is $\Theta = \{r, s, n\}$. This frame may be coarsened into $\Theta' = \{n, \overline{n}\}$ if we just want to know whether the weather is normal or not. The coarsening combines fine elements r and s into a coarse element $\overline{n}$. Thus, we call Θ' a coarsening of Θ or Θ a refinement of Θ'. A more refined frame is able to represent more details than its coarsenings and so a proposition discerned by a coarsening is also discerned by a refinement. The converse is not true.

Each coarse element in a coarse frame maps to a subset of fine elements in a refined frame. If a belief function Bel is defined on a refined frame Θ, it can be carried over to a coarse frame Θ' as a *marginal belief function* as follows. A focal element of the marginal is a set of coarse elements that map to subsets, all of which intersect with the same set of focal elements of Bel. The basic probability number is the sum of the corresponding basic probability numbers of the intersecting focal elements. On the other hand, if a belief function Bel' is defined on a coarse frame Θ', it can also be carried over to a refined frame Θ by using the same probability numbers but replacing each focal element by the union of corresponding mapped subsets. The resulting belief function is called a *vacuous extension*.

Both vacuous extension and marginalization can be easily expressed in the special case when a refined frame is the Cartesian product of two or more independent frames [22]. Suppose $\Theta_1, \Theta_2, ...$ are independent frames. Let I be a set of indices. Then $\Theta(I) = \prod\{\Theta_i \mid i \in I\}$ will be a refinement for all Θ_i $(i \in I)$ so that each element $\theta_i \in \Theta_i$ maps to subset $\{\theta_i\} \times \Theta(I - \{i\})$ in Θ. Given any belief function on $\Theta(I)$ with a mass function $m(A)$, its marginal on $\Theta(J)$, $J \subset I$, is a belief function with mass function $m^{\downarrow J}$: for any $B \subseteq \Theta(J)$,

$$m^{\downarrow J}(B) = \sum\{m(A) \mid A \cap (B \times \Theta(I - J) \neq \phi\}. \tag{13}$$

On the other hand, if a belief function with mass function $m(A)$ is defined on $\Theta(J)$, its vacuous extension to $\Theta(I)$ $(I \supset J)$ is a belief function with mass function $m^{\uparrow I}$: for any $B \subseteq \Theta(J)$,

$$m^{\uparrow I}(B \times \Theta(I - J)) = m(B). \tag{14}$$

It is easy to see that if a belief function Bel is a separable support function, its vacuous extension will also be separable. However, the converse is not true, i.e., the marginal of a separable support function may not be separable. For this reason, Shafer calls such a marginal belief function a *support function*. So a belief function is a support function if it can be extended to a separable

support function. Since any frame is a refinement of itself, a separable support function is itself a support function. Thus, we have four nested classes of belief functions:

$$\left\{\begin{array}{c} simple \\ support \\ functions \end{array}\right\} \subset \left\{\begin{array}{c} separable \\ support \\ functions \end{array}\right\} \subset \left\{\begin{array}{c} support \\ functions \end{array}\right\} \subset \left\{\begin{array}{c} belief \\ functions \end{array}\right\}.$$

As it turns out, a belief function is a support function if and only if the union of all of its focal elements is also a focal element. Thus, not all belief functions are support functions. Moreover, not all support functions are separable. For example, assume $m(\{r, n\}) = 0.2$, $m(\{s, n\}) = 0.5$, and $m(\Theta) = 0.3$. This is a support function since $\{r, n\} \cup \{s, n\} \cup \Theta = \Theta$ is a focal element. However, this is not a separable support function since $\{r, n\} \cap \{s, n\} = \{n\}$ is not a focal element.

3 A Brief History of Concepts

Einstein once said [14], "...creating a new theory is not like destroying an old barn and erecting a skyscraper in its place. It it rather like climbing a mountain, gaining new and wider views, discovering unexpected connections between our starting point and its rich environment." The theory of belief functions arose first from Dempster's attempt in understanding and perfecting Fisher's fiducial approach to probability inference and then from Shafer's elaboration of Dempster's work toward a general theory of reasoning based on evidence.

In the 1960s, due to the work of Leonard J. Savage [32], Bayesian statistics was showing renewed vigor and gaining popularity but, at the same time, was in growing conflict with a school of thought led by Ronald A. Fisher and, increasingly, Jerzy Neyman.

The general statistical inference problem is that, given a sample observation x from a parametric distribution $f(x, \theta)$ with parameter θ, how one could obtain a probability distribution of θ. When reduced to its mathematical essentials, Bayesian inference means starting with a prior probability distribution $p(\theta)$, observing the value x, and computing the conditional distribution of θ given x using Bayes theorem:

$$p(\theta \mid x) = \frac{p(\theta) f(x, \theta)}{\int p(\theta) f(x, \theta) d\theta}. \tag{15}$$

In theory, there is nothing wrong with this formulation. In practice, however, one often finds the conception of prior probabilities vague, arbitrary, or controversial, lacking the spirit of objectivity required by a scientific method.

To overcome the difficulty with prior probabilities, Fisher announced the possibility of obtaining posterior distributions with no need for priors

(see [17]), and called his method the *fiducial argument* to emphasize its differences from the Bayesian argument. In the nutshell, assume $F(x, \theta)$ is a parametric cumulative distribution. Besides x and θ, the fiducial method introduces a so-called *pivotal variable* u, which is assumed to follow the uniform distribution $U(0, 1)$, so that

$$u = F(x, \theta). \tag{16}$$

Suppose, for each value x, $F(x, \theta)$ is monotonic in θ. Equation 16 will admit a unique solution

$$\theta = \theta(u, x) \tag{17}$$

for each $u \in (0, 1)$. Assuming no prior probabilities, Fisher defined the fiducial distribution of θ, given the observed value x, as the distribution of θ implied by (17) when x is regarded as fixed and u is uniformly distributed.

The fiducial method was poorly understood and often led to inconsistencies [6]. The concept of pivotal variables was highly confusing, restrictive, and controversial [7]. Dempster devoted much of his early research career at Harvard to clarifying, extending, and perfecting the method. For example, he once proposed the concept of *direct probabilities* as his interpretation of the fiducial argument [5]. First, to make the derivation of fiducial probabilities explicit, he introduced an arbitrary function $v = V(x)$ so that it along with (16) implied a smooth one-to-one function from x and θ to u and v, and therefore ensured the existence of the following Jacobian:

$$\left| \frac{\partial(u, v)}{\partial(x, \theta)} \right|. \tag{18}$$

Second, in addition to Fisher's assumption that u is uniform in $(0, 1)$, he assumed that v follows an arbitrary distribution $p(v)$ and u is independent of x (and so of v) so that the joint density function of u and v is $p(v)$. Finally, according to the Jacobian formula, the joint density function of x and θ is $p(V(x))$ multiplied by the Jacobian in (18). From this joint distribution, of course, one can compute the conditional probability distribution of θ given x, which is the fiducial (or direct) probability distribution.

Like Bayesian priors, functions $V(x)$ and $P(v)$ are arbitrary and meant to compose a joint distribution, from which a conditional distribution can be obtained. Although $P(v)$ does not enter the final result, a fiducial distribution is generally not free from the choice of $V(x)$. In fact, as Dempster showed, it is independent of $V(x)$ if and only if $F(x, \theta)$ can be transformed into a location parameter family.

The direct probability method did not fully demystify the fiducial argument. Although it explicated the process of deriving fiducial probabilities, it left the concept of pivotal variables unexplained. Some regard the uniform distribution $U(0, 1)$ as analogous to a Bayesian prior. Most importantly, like

the fiducial argument, the method works only if there exists a smooth one-to-one mapping Γ: $u \to \theta$ so that a probability measure for u can be carried to θ by the familiar Jacobian formula.

A breakthrough led to a new theory that unified Bayesian and fiducial arguments. It was first exposited in a paper published in 1966 [8] and republished here as Chap. 2. In this paper, Dempster abandoned Fisher's controversial pivotal variable and replaced it with the concept of a population. Instead of considering u as a pivotal variable, uniformly distributed in $(0, 1)$, he construed u to be a sample individual randomly drawn from a population with probability measure m governing the random sampling operation. Here, m is not necessarily a uniform distribution as in the case of the fiducial argument. Second, instead of (16), Dempster proposed a new model for constructing the mapping from u to θ as follows. Assume each sample individual u corresponds to an observable characteristic x. Assume further that the probability measure m for u induces a probability distribution $f(x, \theta)$ for x with an unknown parameter θ. Thus, one may construct a mapping $u \to x \times \theta$. When the observation x is fixed, it determines a conditional mapping Γ: $u \to \theta$, from which m induces a probability distribution for θ. Interestingly, when Γ is multivalued, the induced distribution for θ is no longer unique. Instead, Γ carries a unique probability measure m to a system of upper and lower probabilities for θ.

Chapter 2 was a milestone, representing not only an advancement of the fiducial argument but also the inception of the idea for a new theory of belief functions. At this point, the basic concepts had already emerged, including the basic probability assignment m, the multivalued mapping Γ, and a device for deriving upper and lower probabilities from m. In Chap. 3, first published in 1967 [9], Dempster abstracted these concepts from the fiducial argument, envisioning a fundamental method of reasoning with imprecise probabilities, based on the idea of obtaining a degree of belief for one event from probabilities for related events. He proposed a general model (S, m, Γ, T) for such reasoning, where S is a source space, m is a probability measure over S, T is a target space, and Γ is a mapping from S to T. If Γ is a one-to-one or many-to-one mapping, it is well known that the probability measure m carries over to T as $p(t) = \sum\{m(s) \mid t = \Gamma(s)\}$. In mathematical essence, Chap. 3 extended the familiar result to the case when Γ is a one-to-many or many-to-many mapping and derived a system of upper and lower probabilities for T based on a probability measure m. Its real thrust, of course, is to view a probability measure as defining degrees of belief, which quantifies a state of partial knowledge arising from a source of imprecise information. Since information is imprecise, it does not always pinpoint a unique value of the variable of interest. Thus, a multivalued mapping is a necessary representation for imprecise information. Since there may be multiple independent sources of information, a mechanism for combining such sources becomes a necessity for a general calculus oriented toward statistical inference and probabilistic reasoning. Therefore, besides the formal definitions of upper and lower probabilities, distributions, and expectations, Chap. 3 presented a rule for deriving upper and lower conditional

probabilities and further generalized it into a rule of combining independent sources of information, which was later called Dempster's rule of combination by Shafer [34].

The concept of upper and lower probabilities can be traced back to Boole [2]. Before Dempster, there were already other approaches to the concept [16, 18, 38, 39]. Dempster's multivalued mappings provides a rigorous device for generating these probabilities. As Chap. 3 showed, however, Dempster's concept is not the same as alternative ones. For example, the set of probabilities compatible with Dempster's upper and lower probabilities is smaller than alternatives. The unique feature of Dempster's concept is to map upper and lower probabilities to a single probability measure, allowing for a more rigorous logic for defining conditioning. The resulting upper and lower conditionals are, of course, not same as upper and lower bonds of conditionals. Using standard notations, let $Bel(A)$ and $Pl(A)$ be Dempster's lower and upper probabilities. Then, given a subset E with $Pl(E) > 0$, Dempster's conditional is

$$Bel(A \mid E) = \frac{Bel(A \cup \overline{E}) - Bel(\overline{E})}{1 - Bel(\overline{E})}. \tag{19}$$

In contrast, let $\mathbf{P}$ be the set of probability measures compatible with Bel: $\mathbf{P} = \{P \mid P(A) \geq Bel(A)\}$. Given E with $P(E) > 0$, we can take Bayesian conditioning of P in $\mathbf{P}$: $P_E(A) = P(A)/P(E)$. Let $\mathbf{P}_E$ be the set of resulting conditionals: $\mathbf{P}_E = \{P_E \mid P \in \mathbf{P}\}$. Then, the lower envelope of $\mathbf{P}_E$ exists when $Bel(E) > 0$: $\forall A \subset E$,

$$\underline{P}(A \mid E) = \frac{Bel(A)}{Bel(A) + 1 - Bel(A \cup \overline{E})}. \tag{20}$$

In general, we have $Bel(A \mid E) \geq \underline{P}(A \mid E)$. Therefore, Chap. 3 not only established the mathematical foundation for the theory of belief functions but also clarified many confusions that later arose in the literature [29]. In fact, to avoid these confusions, Shafer [34] renamed Dempster's upper and lower probabilities into respectively plausibility and belief functions.

Being a statistician, Dempster first explicitly applied his rule of combination to statistical inference. He did this in Chap. 4, first published in 1968 [10]. Although Chap. 4 derived upper and lower probabilities for the same parameters, it did so without explicitly invoking the rule of combination. Chap. 4 framed the inference problem using a formal model (S, m, Γ, T), where S is a population, m is a probability measure governing how each individual may be sampled from the population, and $T = X \times \Theta$ is the product of the set of all possible observations x and the set of all possible parameter values θ. A multivalued mapping $\Gamma : S \to T$ was then used to derive a restricted mapping $\Gamma_\theta : S \to X$ when the parameter θ is fixed or $\Gamma_x : S \to \Theta$ when an observation x is made. Thus, one could obtain two restricted models: (S, m, Γ_θ, X) and (S, m, Γ_x, Θ). The former may be used to derive upper and lower probability for future observations x, and the latter to derive the same for the parameter θ. When there are multiple independent observations, one can produce

one restricted model for each observation and then combine these models using Dempster's rule to derive combined upper and lower probabilities for θ. When a prior distribution $p(\theta)$ is available, it can be regarded as yet another restricted model (Θ, p, I, Θ), where I is the identity mapping, which can be also combined with the restricted models based on sample observations. Therefore, Chap. 4 consolidated the fiducial arguments and Bayesian inference and brought them under the same umbrella of belief functions. It not only showed the feasibility of probabilistic inference without priors but also re-expressed Bayesian inference as the combination of independent sources of information, including priors and sample observations.

As side products of its application of belief functions, Chap. 4 made additional theoretical contributions. The first was the concept of total ignorance and its representation via upper and lower probabilities. This provided a simple resolution to the old controversy about the representation of ignorance via a probability distribution, and led to the concept of vacuous belief function [34] that showcased the superiority of belief functions for subjective judgments. The second was the idea of viewing prior knowledge as a source of information similar to other sources such as sample observations to be combined via Dempster's rule. This idea led to the concept of Bayesian belief functions [34] and embraced Bayesian probabilities as a special case of belief functions. The third was the idea of viewing a multivalued mapping as a random set. This idea led not only to an alternative formalization of the theory of belief functions but also to an alternative perspective on belief functions as the extension of probability distributions over random variables. It also allowed for a rigorous mathematical foundation for belief functions.

Chapters 2–4 established most of the basic ideas and concepts for a new theory of probable reasoning. Without extensions, refinements, and reinterpretations by Glenn Shafer, however, these elements would still have been in the narrow statistical confines of random sampling. While studying for his Ph.D. at Harvard, Shafer got acquainted with Dempster's work. Later he was asked to make a presentation on Dempster's upper and lower probabilities at Princeton. His book—*A Mathematical Theory of Evidence* [34]—was a result of his ensuing effort. Characterized by Shafer's intellectual boldness, the book announced the establishment of a new mathematical theory for probable reasoning as a genuine generalization of or superior alternative to subjective Bayesian theory. To distinguish the theory from theories of imprecise probability, the book renamed Dempster's lower and upper probabilities respectively as belief and plausibility functions. Whereas Dempster had emphasized the derivation of lower and upper probabilities from S, m and Γ, Shafer regarded belief functions as a fundamental concept—an alternative to subjective probabilities. Following Andrei Kolmogorov, who built probability theory on three mathematical axioms, Shafer built his theory of belief functions on three similar axioms, with the additivity of probabilities being replaced by the super-additivity of belief functions. He showed that a belief function satisfied the three axioms if and only if it was as derived from a basic probability

assignment $m(2)$. It was this connection that allowed Shafer to simplify Dempster's four-element model (S, m, Γ, T) into a two element model (m, T), which assigned probabilities m directly to subsets of the target space T while keeping S and Γ implicit. It was also this connection that allowed belief functions to express partial beliefs for probable reasoning using the two basic ideas due to Arthur Dempster: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question (evidence), and Dempster's rule for combining degrees of belief when they were based on independent items of evidence.

Besides providing new terminologies, notations, and the axiomatization, Shafer also greatly extended Dempster's mathematical results. Most notable are the concepts of support functions and weights of evidence. These concepts served two purposes. First, they showed how weights of evidence might be converted into degrees of belief and combined using Dempster's rule, and thus showed how the theory of belief functions could be rebuilt and applied around these concepts. Second, they justified the theory of belief functions from works of Jakob Bernoulli and other ancient scholars on probabilities. It was probably from these works Shafer generated his idea of re-interpreting Dempster's work as a theory of probable reasoning through the combination of evidence.

The notion of weights of evidence can be traced back to Jakob (James, Jacques) Bernoulli in his book *Ars Conjectandi*. Jakob Bernoulli died in 1705. His book was given to the printer by his nephew Nicholas Bernoulli, under the pressure of mathematicians. After it was published in 1713 by the Thurneysen Brothers Press in Basel, *Ars Conjectandi* became the founding document of mathematical probability, replacing *Calculating in Games of Chance* by Christian Huygens, which was the first ever printed book on probability and served as the standard text for over 50 years after 1657. *Ars Conjectandi* consisted of four parts. Part 1 was an improved version of Huygens' book on games of chance with annotations. This part made many well-known contributions in elementary probability theory. For example, the notion of Bernoulli trials, the multiplication rule for independent events, and the Bernoulli distribution were all presented in this part. Part 2 offered a thorough treatment of the mathematics of combinations and permutations, including the numbers known as "Bernoulli numbers." Part 3 solved some complicated problems of games of chance using combinatorics. The final part manifested Bernoulli's crowning achievement in mathematical probability. For example, he proved what we now know as the weak law of large numbers. A complete English translation of the book was done only recently by Edith Dudley Sylla [1]. Part 4 was translated into English by Bing Sung [41] with a preface by Arthur Dempster.

Part 4 of *Ars Conjectandi* envisioned the application of probability theory to economics, morality, and politics. Bernoulli did not in fact make such practical applications. But he did succeed in formulating a concept of mathematical probability that went beyond the application to games of chance. He

characterized probability as a degree of certainty that differs from absolute certainty as a part differs from a whole. The art of conjecture was to measure as exactly as possible the probabilities of things. With respect to games of chance, the symmetry of physical devices suggested we could calculate the probability of a specified outcome as the number of favorable cases divided by the total number of cases. In many other situations, however, such symmetry could not be relied upon and the classical procedure could not be applied. Thus, probability was a measure of imperfect knowledge and was personal in the sense that it varied from person to person according to his knowledge. This statement has credited Bernoulli today as the father of subjective probability theory. Nevertheless, it is instructive to compare Bernoulli's notion with several distinct modern ones of subjective probability. In the personalist theory of Bruno de Finetti, Frank P. Ramsey, and Leonard J. Savage, probabilities may be unknown only insofar as one "fails to know one's own mind" and are measured by the betting ratio at which the person in question is willing to bet on the truth of the statement. In the logical theory of John M. Keynes and Harold Jeffreys, probabilities may be unknown by failure to do logic but no experiment will help check up on logical probability. In the subjective theory of Werner Heisenberg, probability contains the objective element of tendency and the subjective element of incomplete knowledge. An observation cannot predict a result with certainty; what can be predicted is the probability of a certain result, and this probability can be checked by repeating the experiment many times. In contrast, Bernoulli believed that everything was governed by God and causal mechanism. As long as we knew the causes, what could seem to be to one person at one time an uncertain event might be at another time to another person (indeed, to the very same person) a deterministic event. From this comparison, Hacking [19] concluded that Bernoulli's subjectivism was less like the personalist or logical point of view, and more like that of the physicists.

Because he wanted to measure probabilities, Bernoulli was concerned with how to combine evidence of different sorts. He stated that probabilities are estimated by the number of cases and the weight of evidence.[1] His first scheme of combination followed the Port Royal logic of Pascal and distinguished internal versus external evidence. *Internal evidence* arises from the topics—cause, effect, subject, sign, circumstance, or anything that directly connected to the question of interest. *External evidence* appeals to human authority or testimony. His second scheme descended from Gottfried Wilhelm Leibniz's notion of pure and mixed evidence. *Pure evidence* proves a thing with a certain probability without giving a positive probability to the opposite thing, whereas *mixed evidence* proves a thing with a certain probability and proved the opposite with the complementary probability. That Gracchus turned pale

[1] Bernoulli used *argumentum* instead of evidence. This Latin word has a broad sense emcompassing the meanings of the modern English words "evidence" and "argument."

when interrogated is an example of pure evidence. To assess the probability of a thing, one can list all pieces of evidence. If all pieces are mixed, then the probability is the number of favorable cases divided by the total number of cases. The resulting probabilities are additive and complementary. However, if all or some pieces of evidence are pure, Bernoulli formulated a rule of combination, which Shafer [34] showed was a special case of Dempster's rule. The resulting probabilities may not be additive and complementary.

Clearly, Bernoulli's notion of non-additive probabilities was the ancestor of what we now call belief functions. This explains why Shafer reinterpreted lower probabilities as epistemic probabilities or degrees of belief while abandoning the term of lower probability, which can arise as lower bounds over classes of Bayesian probabilities. It was also clearly Bernoulli's idea of probability assessment through combining weights of evidence that motivated Shafer to recast Dempster's theory of random sampling into a theory of evidence and to represent evidence using support functions.

4 Classic Works

Although they are presented chronologically, the classic contributions in this volume can be grouped, at least roughly, by their content and emphasis into seven categories: conceptual foundations, philosophical perspectives, theoretical extensions, alternative interpretations, and applications to artificial intelligence, decision making, and statistical inference.

4.1 Conceptual Foundations

Four chapters may be said to have established the conceptual foundation of belief functions presented in Shafer's book [34]: Chaps. 2–4 by Dempster and Chap. 7 by Shafer.

The previous section has given a detailed account of Chaps. 2–4. In brief, Chap. 2 proposed the multivalued mapping approach to deriving upper and lower probabilities to replace posterior distributions in the absence of Bayesian priors. It was the first belief-function treatment of Fisher's fiducial method. Chapter 3 envisioned the problem of obtaining degrees of belief for one question from a probability measure of a related question through a multivalued mapping. It introduced Dempster's role of combination and a corresponding notion of commonality functions. Chapter 4 explicitly applied Dempster's rule to statistical inference and marked the birth of generalized Bayesian theory or a theory of belief functions.

Chapter 7 extended the concept of belief functions defined in [34] to continuous frames of discernment. Following the approach by Gustave Choquet [4], the chapter considered a subset in a continuous frame as the limit of a sequence of finite subsets, and proposed the concepts of continuity and

condensability. Continuity was defined in the same way as the continuity of a Lebesgue measure. Condensability was a key assumption for the extension: a belief function is *condensable* if its plausibility function Pl satisfies

$$Pl(A) = \sup\{Pl(B) \mid B \subset A \text{ and } B \text{ is finite}\}.$$

The chapter then showed how to extend a continuous or condensable belief function on an algebra of (finite) subsets of Θ—a set of subsets that is closed under both set union and complement operations—to a continuous or condensable belief function on the power set 2^{Θ}. The main tool used for such an extension was Choquet's integral representation theorem, which implies that every belief function can be represented by an *allocation of probability*. Technically, for every belief function Bel on an algebra of subsets of Θ, there exists a homomorphic mapping ρ into a probability algebra with a positive and additive probability measure m such that $\rho(A \cap B) = \rho(A) \cap \rho(B)$ and $Bel(A) = \int \rho(A) dm(\rho)$.

4.2 Philosophical Perspectives

We place in this group Chaps. 6 and 9 by Glenn Shafer, Chap. 13 by Glenn Shafer and Amos Tversky, and Chap. 30 by Arthur P. Dempster. All these chapters justify the theory of belief functions from broader perspectives.

Chapter 6 provided a historical account of non-additive probabilities as well as rules of combining evidence. It focused on the work of Jakob Bernoulli and its extension by Johann Heinrich Lambert, a 18th century scholar. It related these ancient concepts of non-additive probabilities to the modern concept of belief functions and showed that both Bernoulli and Lambert's rules of combination are special cases of Dempster's rule.

Chapter 9 systematically examined the critiques by Bayesian or imprecise probability theorists. Both Bayesian and lower probability theories can appeal to the betting interpretation or the Dutch-Book argument for the semantics of its degrees of belief. What is the semantics of belief for a belief function? In the literature, some authors appeal to the probability of provability [28, 31, 37] or the support of arguments [23, 21]. Nevertheless, Chap. 9 argued that Bayesian, imprecise probability, and belief functions are all constructive theories for probability judgment. They need not rely for their meaning and justification on any behavioral interpretation. Instead, the degree of belief is the result of comparing evidence to knowledge about chances governing the truth. The chapter proposed the randomly coded message as a scale for such a comparison: suppose someone chose a code at random from a list of codes and we knew the probability of each code being chosen. Then $m(A)$ is the sum of probabilities of codes, by which the decoded message is A.

Furthering the idea of constructive probability, Chap. 13 dealt with human judgments of probabilities and belief functions. It illustrated that both Bayesian theory and the theory of belief functions were formal languages for

one to analyze evidence and express his degrees of belief; they had the usual components of a language, including vocabularies, semantics, and syntax. It suggested that making a probability judgment was a process of conducting a mental experiment and hence the quality of the experimental design affected the quality of the judgment. The chapter offered some alternative designs for using the languages of Bayesian probabilities and belief functions. For example, the total-evidence design often used with Bayesian theory is distinguished from the belief function that emphasizes the decomposition of evidence. The chapter emphasized that theories of subjective probability (including belief functions) were not psychological models, either normative or descriptive, for making judgments. An experimental design for using such a theory (or its semantics and syntax) must guide the process of making probabilistic judgments.

Chapter 30 is a new contribution based on the 1998 R.A. Fisher Memorial Lecture.[2] The theory of belief functions arose from the need for a new scientific method unifying various statistical methods, including fiducial and Bayesian methods. As opposed to Bayesian, Fisherian, or frequentist statistics, Dempster proposed logicist statistics as a unified way to study principled and explicit reasoning about uncertainty. The key concept was formal subjective probability, which interprets each numerical probability as a degree of certainty reflecting specific formalized evidence and information within a formal mathematical model. Dempster showed that this concept encompasses both modern Bayesian and traditional Fisherian thinking, and he interpreted frequentist theory in a way that gives appropriate weights to both science and mathematics, and to both subjective and objective elements. He also suggested that the Dempster-Shafer theory embodies a more suitable paradigm for logicist statistical inference than Bayesian inference and is logicist in a fundamental way because it integrates nonprobabilistic "propositional" logic with probabilistic reasoning.

4.3 Theoretical Extensions

This group contains Chap. 5 by Hung T. Nguyen, Chap. 11 by Ronald R. Yager, Chap. 15 by Nevin L. Zhang, Chap. 19 by Alain Chateauneuf and Jean-Yves Jaffray, and Chap. 21 by John Yen. These five chapters extended the theory of belief functions in various ways.

Chapter 5, by Nguyen, was the first research work on belief functions published by someone other than Dempster and Shafer. It carried out the idea in Chap. 3 by Dempster that a multivalued mapping might be considered a random set and established the connection between belief functions and random sets. It showed that, in finite cases, the probability distribution

[2] The Fisher Lectureship and Award was established in 1963 by the Committee of Presidents of Statistical Societies to recognize the importance of statistical methods for scientific investigations.

of a random set is a basic probability assignment and a belief function is deduced from the probability distribution of the random set. It characterized the condensability of belief functions of Chap. 7 using the notion of regularity of probability measures. It showed that a plausibility function is condensable if and only if the corresponding probability distribution of a random set is regular.

In probability theory, entropy is a measure of the disorder and randomness present in a distribution. In fuzzy logic, specificity is an overall measure of how much a possibility distribution points to one and only one element as the manifestation of a fuzzy variable. A belief function has both randomness and non-specificity components. Thus, Chap. 11, by Yager, developed similar concepts for belief functions. For a belief function with mass function m and plausibility function Pl, its entropy is

$$E = -\sum\{m(A)\log(Pl(A)) \mid A \subseteq \Theta\}.$$

This formula reduces to Shannon entropy for Bayesian belief functions. It attains zero entropy for consonant belief functions and the maximum entropy when focal elements are disjoint and when the belief mass is equally distributed among all focal elements. The specificity of a belief function with mass function m is defined as

$$S = \sum\{\frac{m(A)}{|A|} \mid \phi \neq A \subseteq \Theta\}.$$

This measure reduces to the specificity of a fuzzy variable for a consonant belief function. It reaches the minimum value for a vacuous belief function and the maximum value for Bayesian belief functions. Chapter 11 led to many studies on the measurement of total uncertainty encompassing both randomness and nonspecificity. One noteworthy contribution [27] uses a set of reasonable axioms to derive measures such as

$$H = \sum\{m(A)\log(\frac{|A|}{m(A)}) \mid \phi \neq A \subseteq \Theta\}.$$

This measure has many desirable features, including additivity for independent belief functions and reduced computational complexity.

Chapter 15, by Zhang, was one of few contributions that directly improved the classic book by Shafer [34]. Note that the weight of evidence provides a full assessment of evidence for simple and separable support functions. Can a similar concept be extended to support functions that may not be separable? Shafer [34] approached the problem indirectly through notions of internal conflict and impingement. For any separable support function T, let W_T and v_T be respectively its weight of internal conflict and impingement function. For any support function S over Θ, let ϵ_S be the set of all its extensions that are separable support functions over some refinements of Θ. Then the

weight of internal conflict for S was defined as the minimum weight of internal conflict among all separable support functions in ϵ_S:

$$W = \inf\{W_T \mid T \in \epsilon_S\}. \tag{21}$$

Similarly, the impingement function of S was derived from those of all its separable extensions: for any subset $A \subset \Theta$,

$$v(A) = \inf\{v_T(\omega(A)) \mid T \in \epsilon_S,\ \omega \textit{ is a refinement mapping}\}. \tag{22}$$

Since a separable support function itself is a support function, the definitions in (21) and (22) should also apply to separable support functions and the result should be consistent, i.e., if S is a separable support function, then $W = W_S$ and $v = v_S$. Shafer [34] proved the consistency by assuming the weight-of-conflict conjecture, which has not been proved to be true yet. This chapter proved the consistency without the conjecture.

As we see in Chaps. 5 and 7, a belief function is a monotone capacity of infinite order whereas a mass function is the Möbius inversion of the capacity. Chapter 19, by Chateauneuf and Jaffray, studied the properties of capacities of all orders, whose relationship is that, for any $K \geq 2$, if a capacity is K-monotone, then it is also L-monotone for $K \geq L \geq 2$ and 1-monotone (or monotonic in usual sense) if $f(\theta) \geq 0$ for any $\theta \in \Theta$. A capacity is defined as ∞-monotone if it is K-monotone for any $K \geq 2$. The chapter obtained some useful results characterizing the capacities through Möbius transformations. For example, it showed that, capacity f is K-monotone ($K \geq 2$) if and only if, for any A and $C \subset \Theta$ with $2 \leq |C| \leq K$, its Möbius inversion m satisfies:

$$\sum_{C \subset B \subset A} m(B) \geq 0.$$

The chapter also characterized probability distributions that dominate (or "are compatible with" in terms of Chap. 2) a belief function. It showed that if the probability distribution P satisfies $P(A) \geq f(A)$ for any A, then P is the weighted average of the Möbius inversions of f:

$$P(x) = \sum_{x \in B} \lambda(B, x) m(B).$$

It generalized a result in Chap. 3 by Dempster and showed f is ∞-monotone if and only if every probability distribution dominating f is the weighted average of Möbius inversions.

Many scholars in the area of fuzzy logic consider Chap. 21, by Yen, an outstanding paper. It is a favorite reference on the fuzzification of belief functions. It studied the computation of beliefs and plausibilities for fuzzy sets and extended Dempster's rule to fuzzy logic. It significantly improved other approaches by Zadeh, Ishizuka, Yager, and Ogawa while maintaining the semantics of the Dempster-Shafer theory of belief functions as well as

possibility theory. It brought together belief functions and fuzzy logic into a hybrid approach to reasoning under various kinds of uncertainty in intelligent systems. The chapter started with a novel viewpoint, from which the computation of $Bel(A)$ was formulated as a linear programming problem:

$$\min \sum_{x \in A} \sum_{B} m(x, B)$$
$$s.t. \;\; m(x, B) \geq 0; m(x, B) = 0 \; \forall x \notin B; \sum_{x} m(x, B) = m(B),$$

here $m(x, B)$ denoted the probability mass allocated to x from $m(B)$. Then, when A was a fuzzy set, the chapter proposed to extend the problem into one of minimizing the extended objective function:

$$\sum_{x \in A} \sum_{B} m(x, B) \mu_A(x),$$

here $\mu_A(x)$ denoted the membership of x in A. If all focal elements were crisp (non-fuzzy), then the solution to the generalized problem is

$$Bel(A) = \sum m(B) \inf_{x \in B} \mu_A(x).$$

If any focal element B is fuzzy, it will be broken into one or more crisp focal elements, each of which is an $\alpha - cut$ of B:

$$B_\alpha = \{x \mid x \in B, \; \mu_B(x) \geq \alpha\},$$

with a basic probability mass

$$m(B_\alpha) = (\alpha_i - \alpha_{i-1}) m(B),$$

here $\alpha_0, \alpha_1, \alpha_2, ..., \alpha_n$ is a series of membership degrees of increasing order with $\alpha_0 = 0$ and $\alpha_n = 1$. For example, if focal element $B = \{(yound, 0.4), (old, 0.7)\}$ with $m(B) = 0.8$, then we get two α–$cuts$ as follows: $B_{0.4} = \{yound, old\}$ and $B_{0.7} = \{old\}$ with basic probability masses $m(B_{0.4}) = (0.4 - 0) \times m(B) = 0.32$ and $m(B_{0.7}) = (0.7 - 0.4) \times m(B) = 0.24$. Then, $Bel(A)$, for any fuzzy set A, is

$$Bel(A) = \sum_{B} m(B) \sum_{i} (\alpha_i - \alpha_{i-1}) \inf_{x \in A_{\alpha_i}} \mu_A(x).$$

The approach to extending Dempster's rule was also novel. It considered a multivalued mapping $S \rightarrow T$ as a compatibility relation $S \times T$ and generalized it to a fuzzy relation $C : 2^{S \times T} \rightarrow [0, 1]$, which is a joint possibility distribution. It considered Dempster's rule as the combination of compatibility relations and generalized it as the combination of fuzzy relations, which in turn is equivalent to the multiplication of noninteractive possibility distributions.

This led to the generalized rule for combining fuzzy belief functions. Let m_1 and m_2 be two fuzzy mass functions. Then,

$$m_1 \oplus m_2(C) = \frac{\sum_{A \cap B = C} \max_x \mu_{A \cap B}(x) m_1(A) m_2(B)}{1 - \sum_{A,B} (1 - \max_x \mu_{A \cap B}(x)) m_1(A) m_2(B)}.$$

4.4 Artificial Intelligence

Five chapters apply belief functions to uncertain reasoning in artificial intelligence. Chapter 8 by Jeffrey Barnett was the first paper dealing with computational issues in implementing Dempster's rule of combination. It proposed an algorithm based on the very strong assumption that each piece of evidence either confirms or denies a single proposition, i.e., all focal elements are singletons or their negations. Chapter 12 by Jean Gordon and Edward Shortliffe proposed an improved algorithm capable of handling hierarchical evidence, where focal elements and their negations could be arranged in a tree-like structure. To avoid the exponential explosion in computations, the algorithm employed approximation to combine evidence. The approximation was usually reasonable but did give unsatisfactory results in the case of highly conflicting evidence. In addition, the approach did not produce the degrees of belief for all focal elements involved in the computation except for those in the tree. Chapter 18 by Glenn Shafer and Roger Logan presented a further improvement that is at least equally efficient while removing all the above limitations. These chapters built upon each other technically but are all included here because they made the history in distinct ways. Chapter 8 coined the name "Dempster-Shafer theory" and introduced it to the AI community. It was clearly one of the initial sources that led Edward Shortliffe to realize the relevance and applicability of belief functions to the issues addressed by the certainty factor model implemented in the medical advising program MYCIN. Because of their role in MYCIN, Gordon and Shortliffe were probably the most influential of the authors who made belief functions widely known as "the Dempster-Shafer theory" to AI researchers.

Chapter 16 by John D. Lowrence, Thomas D. Garvey, and Thomas M. Strat proposed a formal framework based on belief functions for knowledge representation and uncertainty reasoning in expert systems, setting belief functions up as an alternative to rules, frames, and semantic networks. It introduced the new term "evidential reasoning" for the framework and demonstrated its application in the Gister project at SRI International. Stemming from the application of belief functions to Navy intelligence problems, Chap. 16 was very practical in nature. Its approach to knowledge representation, i.e., modeling compatibility relations, provided a perfect example to illustrate the applicability of belief functions to real problems.

In the framework of Chap. 16, each piece of knowledge is represented by a belief function. Making inferences boils down to combining all component

belief functions and marginalizing the joint belief function into a subframe of discernment (see definition in (13)). Of course, such a straightforward approach would be very inefficient, if not infeasible, when the size of the joint frame is large. A creative solution to the problem is so-called *local computation* that computes marginals without computing the joint. The basic idea is to arrange all the frames of discernment into a tree-structured graph, called *a join-tree* or *Markov tree*, and propagate knowledge by sending and absorbing messages step-by-step in the tree. Each step involves sending a message from a node to a neighbor and thus involves only a small number of frames that are near each other in the join-tree.

Scholars in belief functions, including Glenn Shafer, Prakash P. Shenoy, Augustine Kong, and Khaled Mellouli, pioneered the local computation method. Later they demonstrated the applicability of this method to other calculi, including Bayesian probabilities and fuzzy logics. Chapter 20, by Shenoy and Shafer, presented an abstract framework that covered diverse local computation models as special cases. It characterized many types of computational problems as one of applying two operators: combination and marginalization, where combination corresponds to the integration of two or more factors into a joint model and marginalization corresponds to the projection of a model to a subset of variables. The chapter showed that local computation was applicable to such problems if the two operators satisfied four axioms. For belief functions, for example, these axioms can be represented as follows:

Axiom 4 *Combination operator* $\oplus$ *is commutative: for any* Bel_1 *and* Bel_2,

$$Bel_1 \oplus Bel_2 = Bel_2 \oplus Bel_1.$$

Axiom 5 *Combination operator* $\oplus$ *is associative: for any* Bel_1, Bel_2, *and* Bel_3,

$$Bel_1 \oplus (Bel_2 \oplus Bel_3) = (Bel_1 \oplus Bel_2) \oplus Bel_3.$$

Axiom 6 *Marginalization is consonant: for any Bel on the frame* $\Theta(I)$ *and* $K \subset J \subset I$. *Then*

$$(Bel^{\downarrow J})^{\downarrow K} = Bel^{\downarrow K}.$$

Axiom 7 *Marginalization is distributive over combination: for any* Bel_1 *and* Bel_2 *and* I,

$$(Bel_1 \oplus Bel_2)^{\downarrow I} = (Bel_1)^{\downarrow I} \oplus (Bel_2)^{\downarrow I}.$$

Chapter 20 also presented the Shenoy-Shafer architecture for carrying out local computation over a Markov tree, and demonstrated the algorithm using an example of probability propagation. Compared with other similar approaches (e.g., [24]), this architecture gains some efficiency by avoiding divisions, which are required by other methods for obtaining conditional probabilities.

4.5 Decision Making

The theory of belief functions is not meant to be a normative or descriptive theory for decision making. Thus, it does not provide normative axioms or behavioral predictions on how to make decisions and judgments. Because of its expressive power in encoding evidence or modeling uncertainty, however, it has exceptional prescriptive value as a decision support tool. Here we review four chapters demonstrating creative use of belief functions for the purpose, including Chap. 23 by Rajendra P. Srivastava and Glenn Shafer, Chap. 24 by Ronald R. Yager, Chap. 27 by Galina Rogova, and Chap. 29 by Thierry Denoeux.

Chapter 23, by Srivastava and Shafer, applied belief functions to audit decision-making. The chapter derived analytical expressions of the audit risk at three levels: the financial statement level, the account level, and the audit objective level. It made a distinct contribution to the field by showing how to interpret and use plausibility numbers to encode accounting evidence. It also proposed a hierarchical network for evidential reasoning and dealt with belief propagation through the "AND" gates, which were inherent in business decision problems.

There have been numerous attempts to incorporate belief functions into expected utility theory to take advantage of their flexibility in uncertainty modeling. Chapter 24, by Yager, showcased such attempts. It is included here because it is theoretically sound and computationally feasible. Whereas other work reduced Dempster-Shafer degrees of belief to probabilities for use as decision weights, this chapter proposed deriving decision weights from a mathematical programming model. Once we set a pessimism level—a necessary concept for decision making under uncertainty—entropy maximization problem gives weights to be assigned to each outcome within a focal element. The weights then determine the weighted average value of outcomes in the focal element, which along with the corresponding basic probability numbers determine an overall value for each choice. It is shown that the formalism unifies several common decision models for decision-making under risk, uncertainty, and ignorance. Its ordered weighting mechanism is also consistent with psychological findings that have led decision theorists to generalize *expected utility theory* to so-called *rank-dependent utility* [25, 26, 30].

Chapter 27, by Rogova, is a real application with real results. The topic is very timely. In machine learning, the idea of boosting, i.e., combining simple poor learners to form an ensemble that outperforms individual single ensemble members while avoiding overfitting, is gaining a lot of interest in the last decade. In theory, it is known that learners, each performing only slightly better than random, can be combined to form an arbitrarily good ensemble hypothesis [20]. Schapire [33] was the first to provide a provably polynomial time boosting algorithm. He and his colleagues [13] applied boosting to a real-world optical character recognition by using neural networks as base learners. Chapter 26 demonstrated the application of Dempster's rule to the

same problem. Interestingly, it also used neural networks as base learners. It showed that the proposed approach allowed 15–30% reduction of misclassification error compared to the best individual classifier. The method made Eastman Kodak one of the small group of the leaders in an industrial competition for the best optical recognition algorithm.

Chapter 29, by Denoeux, is considered an outstanding application of belief functions to decision making. It proposed a new approach to pattern classification that considered each of the k-nearest neighbors as an item of evidence and used Dempster's rule of combination to pool all evidence together to form a judgment concerning the class membership of a new incoming pattern. Simulation results showed that the proposed approach outperformed the classic voting k-nearest neighbor approach as well as its distance-weighted variant.

4.6 Statistical Inference

Parametric statistical inference is not only the source of motivation for the theory of belief functions but also one of its most important application domains. Chapter 4 demonstrated the potential of belief functions for unifying the traditional fiducial argument and modern Bayesian inference. Here we review three additional chapters revisiting the problem of parametric inference using belief functions, including Chap. 10 by Glenn Shafer, Chap. 22 by Jean-Yves Jaffray, and Chap. 25 by Philippe Smets.

In his book [34], Shafer suggested translating each observation into a consonant belief function on a parameter based on the normalized likelihood. He recognized that this approach does not possess the desirable property that the result using a set of n independent observations be equal to the combination of the n belief functions obtained from the individual observations. Chapter 10, by Shafer, discussed three alternative approaches, including the fiducial argument, the generalized Bayesian method of Chap. 4, and the conditional embedding method of Chap. 25 (see below). It showed that these methods produce coherent results when the nature of the evidence establishing the parametric model is taken into account.

Chapter 22, by Jaffray, studied the effect of Bayesian conditioning when a belief function is (mis)understood as the lower envelope of compatible probability measures. It obtained two important results. First, it reproved the result by Fagin and Halpern [15] that the lower envelope of all Bayesian conditionals is still a belief function, and going beyond Fagin and Halpern, it developed an explicit expression for the mass function for the lower envelope. Second, it showed that the resulting lower envelope does not characterize the set of all conditionals. Let $\mathbf{Q}_E$ be the set of Bayesian conditionals that dominate $\underline{P}(A \mid E)$ (see (20)). Then, $\mathbf{P}_E \subset \mathbf{Q}_E$ if and only if there exist subsets A and B such that $Bel(A \cap B) > 0$, $Bel(A \cup B) < 1$, and $Bel(A \cup B) > Bel(A) + Bel(B) - Bel(A \cap B)$ (see Sect. 3 for the definition of $\mathbf{P}_E$). Also, $\mathbf{P}_E \subset \mathbf{Q}_E$ if and only if there exist E and F with $F \subset E$ and $Bel(F) > 0$ such that the lower envelopes of Bayesian conditionals do not

satisfy $\underline{P}((A \mid E) \mid F) = \underline{P}(A \mid F)$, which is observed by both Bayesian and Dempster's conditioning.

A belief function $Bel(A)$ may be re-expressed in a conditional form as $Bel(A \mid E)$ given evidence E. Then Dempster's rule may be called the *conjunctive rule of combination*, because $Bel_1(A \mid E_1) \oplus Bel_2(A \mid E_2)$ is the combined belief function when both E_1 and E_2 are true. Chapter 25, by Smets, proposed the *disjunctive rule of combination* that allows the combination of two belief functions induced by two pieces of evidence, of which only one can be true. The disjunctive rule is intuitive when applied to parametric inference problems. Suppose B is a set of possible parameter values, one of which is true. For each $\theta \in B$, let us assume there is a belief function $Bel(A \mid \theta)$ representing the likelihood that the true value of X is in A when the parameter is θ. Then the combination of these belief functions follow the disjunctive rule as

$$Bel(A \mid B) = \prod\nolimits_{\theta \in B} Bel(A \mid \theta).$$

The disjunctive rule corresponds the multiplication of belief functions whereas the conjunctive rule corresponds to the multiplication of commonality functions. Based on the disjunctive rule, the chapter derived the generalized Bayesian theorem, where conditional probabilities are replaced by belief functions and prior probabilities by vacuous belief functions. Let B be a set of possible parameter values and A be a set of observations. Let $Bel(A \mid \theta)$ be the likelihood that X is in A given parameter θ. Then, the generalized Bayesian formula represents the posterior belief of B given A as follows:

$$Bel(B \mid A) = \prod\nolimits_{\theta \in \overline{B}} Bel(\overline{A} \mid \theta) - \prod\nolimits_{\theta \in \Theta} Bel(\overline{A} \mid \theta).$$

Some results in Chap. 25 were initially developed in an unpublished dissertation [36]. The generalized Bayesian theorem permits the induction of a belief function for parameters from an observation, leading to a new statistical method, called *conditional embedding*, which was extensively discussed in Chap. 10. Here the author represented them in his framework of transferable belief functions (see below) and attempted to develop a new approach for belief function propagation in a directed belief network.

4.7 Alternative Interpretations

Besides theoretical foundations, perspectives, advances, and applications, there have been tens of studies targeting alternative formalisms and interpretations of belief functions. Here we review four representative ones: Chap. 14 by Didier Dubois and Henri Prade, Chap. 17 by Enrique H. Ruspini, Chap. 26 by Jürg Kohlas and Paul-André Monney, and Chap. 28 by Philippe Smets and Robert Kennes.

There are many connections between *fuzzy logic* and belief functions. As we have seen earlier, possibility and necessity functions are consonant plausibility and support functions that have nested focal elements. Chapter 14, by

Dubois and Prade, exposed another connection between bodies of evidence and fuzzy sets. The classic concept of a set is simply a collection of elements, e.g., $A = \{x, y, z\}$. The concept of a *fuzzy set* extends it to include a *membership function* $m \rightarrow [0, 1]$ describing a graded assessment of the membership of elements in relation to a set. For example, $A = \{(x, 0.3), (y, 0.7), (z, 1)\}$ is a fuzzy set consisting of elements x, y, and z with membership grades 0.3, 0.7, and 1. Chapter 14 viewed a belief function as a further generalization of fuzzy logic and interpreted a body of evidence to be an extended fuzzy set, where an element was replaced by a focal element and a membership grade was replaced by a basic probability number. For example, $A = \{(\{x, y\}, 0.2), (\{z\}, 0.5), \{x, y, z\}, 0.3)\}$ is a body of evidence representing a belief function with $m(\{x, y\}) = 0.2$, $m(\{z\}) = 0.5$, and $m(\{x, y, z\}) = 0.3$. Chapter 14 studied belief functions using this formalism and introduced the notions of extended set operations such as union, intersection, and complementation to bodies of evidence. It discussed and compared four alternative definitions of set inclusion on bodies of evidence. Since it was easier to deal with consonant plausibility and support functions, the chapter applied the notions of inclusion, and pioneered the research on *possibilistic approximation* of bodies of evidence.

Recall that Chap. 9, by Shafer, interpreted belief functions as a constructive theory for probability judgment, and proposed the randomly coded message as the metaphor for understanding the semantics of belief functions. There was another popular interpretation that understood a degree of belief as the *probability of provability* [37, 29]. Formally, suppose we are given a set of logical theories, each logical theory is characterized by a set of axioms, and each theory is assigned a probability such that the probabilities add up to 1. The belief in a proposition A is then the sum of the probabilities of the theories from which A follows as a logical consequence. Chapter 17, by Ruspini, presented a similar interpretation based on the probabilities of a modal proposition toward developing a formal theoretical foundation for evidential reasoning as proposed by Lowrance, Garvey, and Strat in Chap. 16. In particular, it extended Carnap's notion of the *epistemic universe* [3] by including all possible combined descriptions of not only the state of the real world but also the state of knowledge that certain rational agents have about it. It showed that the probabilities defined over a *sigma algebra* of subsets of the epistemic universe have the properties of belief and mass functions and can represent the effect of evidence on the state of knowledge of the rational agents. The epistemic probabilities also induces lower and upper probabilities in the truth algebra that are identical to the interval bounds derived in Chap. 3. Finally, the chapter applied the epistemic logic approach to the problem of knowledge integration and obtained an *additive combination formula* for integrating a wide variety of knowledge of both dependent and independent sources. Under the assumptions of probabilistic independence, the formula is reduced to Dempster's rule of combination.

Chapter 26, by Kohlas and Monney, presented the theory of hints, another interpretation or formalism of the Dempster-Shafer theory of belief functions based on multivalued mapping Γ from a probability space (Ω) to another space of interest (Θ). As we explained earlier, Dempster's original model was $(\Omega, P, \Gamma, \Theta)$, which in fact was exactly the same as *the model of hints*. The difference lies at the interpretation of Ω, which Fisher called the sample space of a pivotal variable, Dempster called the population of sample individuals, but here Kohlas and Monney called the space of *arguments*. Note that in his axiomatic approach, Shafer made the elements Ω and Γ implicit and assigned basic probability numbers directly to subsets of Θ. Chapter 26 argued that the model of a hint contains more information than its derived belief function does, and allows for a straightforward and logical derivation of Dempster's rule for combining independent and dependent bodies of information.

Chapter 28, by Smets and Kennes, presented the transferable belief model (TBM), a subjectivist and non probabilistic view of the Dempster-Shafer theory of evidence. In response to the need for integrating belief functions into a normative decision theory such as expected utility theory, the TBM distinguished clearly the *credal level*, where beliefs are entertained, from the *decision level* where standard utility theory applies, the belief functions being converted into probabilities using the *pignistic transformation*. Another main idea underlying the TBM is the notion of unnormalized belief function and unnormalized conjunctive rule of combination, and the interpretation of the mass $m(\emptyset)$ assigned to the empty set, under the *open-world assumption*, as a degree of belief in the event that the frame of discernment does not contain the true value of the variable of interest.

5 Conclusion

In this chapter, we reviewed the basic concepts and major results presented in Glenn Shafer's book, provided a brief history of the conceptual development, and summarized the major contributions of the selected classic works.

In this volume we deliberately did not include any papers that involve misunderstandings of basic concepts. This includes well known papers by Lotfi A. Zadeh [43] and Judea Pearl [29]. Zadeh criticized the normalization procedure in Dempster's rule of combination. He used an example to show that, in the case of combining two highly conflicting pieces of evidence, the result is not intuitive, although Shafer thought otherwise [35]. Because of this criticism, many authors introduced the "open world" hypothesis and assigned a non-zero basic probability number $m(\emptyset)$ to the empty set (see Chap. 27). Judea Pearl [29] was mainly concerned with the inability of belief functions to represent imprecise probabilities. This concern was addressed 40 years ago by Dempster (see Chap. 2). A belief function was never meant to replace or represent an imprecise probability, which involves a larger set of compatible probability functions than a belief function does. Instead, it is meant to be

a faithful representation of knowledge based on evidence and to combine the knowledge obtained from multiple independent pieces of evidence for making provable probable inferences.

With respect to future research on belief functions, Dempster [11] called for more realistic applications of belief functions to complex systems. He stressed the critical need for credible and tractable models to represent the details of complex systems where quantified uncertainties cannot be obtainable through more traditional routes. He suggested the development of Fisher pivotals and efficient inference algorithms, in particular two-stage MC and MCMC methods, in conjunction with simplification from local computation with graphical structures. In order to improve public awareness of belief functions, Dempster [12] recently suggested a new semantics whereby every proposition A is associated with a triple (p, q, r), where p is the probability "for" A, i.e., $Bel(A)$, q is the probability "against" A, i.e., $Bel(\overline{A})$, and r is the probability of "don't know", i.e., $Pl(A) - Bel(A)$. He showed how this semantics can coherently interpret the notion of p-value, which is often misconstrued as a Bayesian probability "for" the null hypothesis. Theoretically, open problems still remain. For example, in earlier chapters Dempster left some questions on asymptotic properties of the combined belief function when the number of pieces of evidence approaches infinity. In this chapter, we reviewed Shafer's the weight-of-conflict conjecture that is still unsolved, although Chap. 14 showed that it was not needed for justifying the concepts of weight of internal conflict and impingement for a support function. Another problem is posed by Bayesians who seek behavioral justifications of belief functions. Formally, is there a set of behavioral axioms that justifies the existence of a belief function? In other words, are there any necessary and sufficient conditions in terms of how people make choices or judgments in the face of uncertainty underlying a class of belief functions appropriate for the representation of the uncertainty?

References

[1] Bernoulli, J. *The Art of Conjecturing: Together with His Letter to a Friend on Sets in Court Tennis.* Johns Hopkins University Press, 2006. Translated by Edith Dudley Sylla.

[2] Boole, G. *An Investigation into the Laws of Thought.* Walton and Maberly, London, 1854. Reprinted 1951, Dover, NY.

[3] Carnap, R. *Meaning and Necessity.* University of Chicago Press, Chicago, Illinois, 1956.

[4] Choquet, G. Theory of capacities. *Ann. Inst. Fourier 5* (1953), 131–295.

[5] Dempster, A. P. On direct probabilities. *Journal of the Royal Statistical Society Series B 25* (1962), 100–110.

[6] Dempster, A. P. Further examples of inconsistencies in the fiducial argument. *Annals of Mathematical Statistics 34* (1963), 884–891.

[7] Dempster, A. P. On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc. 59* (1964), 56–66.

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

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

[10] Dempster, A. P. A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B 30* (1968), 205–247.

[11] Dempster, A. P. Belief functions in the 21st century: A statistical perspective. In *Proceedings of Insitute for Operations Research and Management Science Annual Meeting (INFORMS-2001)* (Miami Beach, FL, 2001).

[12] Dempster, A. P. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning* (2006), in press.

[13] Drucker, H., Schapire, R. E., and Simard, P. Y. Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence 7* (1993), 705–719.

[14] Einstein, A., and Infeld, L. *The Evolution of Physics.* Simon and Schuster, New York, 1961.

[15] Fagin, R., and Halpern, J. Y. A new approach to updating beliefs. In *Uncertainty in Artificial Intelligence 6*, P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, Eds. Morgan Kaufmann, San Mateo, CA, 1991, pp. 317–325.

[16] Fishburn, P. *Decision and Value Theory.* Wiley, New York, 1964.

[17] Fisher, R. A. Inverse probability. *Proc. Camb. Phil. Soc. 26* (1930), 154–57, 172–173. Reprinted in Bennett, J. H. (1971). *Collected Papers of R. A. Fisher* 2, Univ. of Adelaide.

[18] Good, I. The measure of a non-measurable set. In *Logic, Methodology and Philosophy of Science*, E. Nagel, P. Suppes, and A. Tarski, Eds. Stanford University Press, Stanford, 1962, pp. 319–329.

[19] Hacking, I. *The Emergence of Probability.* Cambridge University Press, New York, 1975.

[20] Kearns, M., and Valiant, L. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM 41* (1994), 67–95.

[21] Kohlas, J., and Monney, P.-A. *A Mathematical Theory of Hints.* Springer, 1995.

[22] Kong, A. *Multivariate Belief Functions and Graphical Models.* PhD thesis, Department of Statistics, Harvard University, Cambridge, MA, 1986.

[23] Laskey, K. B., and Lehner, P. E. Assumption, belief and probabilities. *Artificial Intelligence 41* (1989), 65–77.

[24] Lauritzen, S. L., and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society Series B 50* (1988), 157–224.

[25] Liu, L. A note on Luce-Fishburn axiomatization of rank-dependent utility. *Journal of Risk and Uncertainty 28*, 1 (2004), 55–71.

[26] Luce, R. D., and Fishburn, P. C. A note on deriving rank-dependent linear utility using additive joint receipts. *Journal of Risk and Uncertainty 11* (1995), 5–16.

[27] Pal, N. R., Bezdek, J. C., and Hemasinha, R. Uncertainty measures for evidential reasoning II: A new measure of total uncertainty. *International Journal of Approximate Reasoning 8* (1993), 1–16.

[28] Pearl, J. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[29] Pearl, J. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning 4* (1990), 363–389.

[30] Quiggin, J. A theory of anticipated utility. *Journal of Economic Behavior and Organization 3* (1982), 323–343.

[31] Ruspini, E. H. The logical foundations of evidential reasoning. Tech. rep., SRI International, Menlo Park, California, 1986.

[32] Savage, L. J. *The Foundations of Statistics*. Wiley, New York, NY, 1954.

[33] Schapire, R. E. The strength of weak learnability. *Machine Learning 5* (1990), 197–227.

[34] Shafer, G. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[35] Shafer, G. Belief functions and possibility measures. In *The Analysis of Fuzzy Information*, J. Bezdek, Ed., vol. 1. CRC Press, Boca Raton, FL, 1987, pp. 51–84.

[36] Smets, P. *Un modle mathmatico-statistique simulant le processus du diagnostic mdical*. PhD thesis, Universit Libre de Bruxelles, Bruxelles, Belgium, 1978.

[37] Smets, P. Probability of provability and belief functions. *Logique et Analyse 133-134* (1993), 177–195.

[38] Smith, C. A. B. Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society Series B 23* (1961), 1–25.

[39] Smith, C. A. B. Personal probability and statistical analysis (with discussion). *Journal of the Royal Statistical Society Series A 128* (1965), 469–499.

[40] Srivastava, R. R., and Shafer, G. Belief-function formulas for audit risk. *The Accounting Review 67*, 2 (1992), 249–283.

[41] Sung, B. *Translations from James Bernoulli (with a preface by A. P. Dempster)*. Department of Statistics, Harvard University, Cambridge, Massachusetts, 1966.

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

[43] Zadeh, L. A. Review of *A Mathematical Theory of Evidence*. *AI Magazine 5* (1984), 81.