6-1-1995

INFERENCE AND DECISION-MAKING WITH PARTIAL KNOWLEDGE

S. Rasoul Safavian
Purdue University School of Electrical and Computer Engineering

David Landgrebe
Purdue University School of Electrical and Computer Engineering

INFERENCE AND DECISION-MAKING WITH PARTIAL KNOWLEDGE

S. RASOUL SAFAVIAN
DAVID LANDGREBE

TR-ECE 95-19
JUNE 1995

SCHOOL OF ELECTRICAL
AND COMPUTER ENGINEERING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1285

INFERENCE AND DECISION-MAKING WITH PARTIAL KNOWLEDGE

S. Rasoul Safavian
David Landgrebe

June, 1995

School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana 47907-1285

TABLE OF CONTENTS

ABSTRACT

Bayesian inference and decision making requires elicitation of prior probabilities and sampling distributions. In many applications such as exploratory data analysis, however, it may not be possible to construct the prior probabilities or the sampling distributions precisely.

The objective of this thesis is to address the issues and provide some solutions to the problem of inference and decision making with imprecise or partially known priors and sampling distributions. More specifically, we will address the following three interrelated problems: (1) how to describe imprecise priors and sampling distributions, (2) how to proceed from approximate priors and sampling distributions to approximate posteriors and posterior related quantities, and (3) how to make decisions with imprecise posterior probabilities.

When the priors and/or sampling distributions are not known precisely, a natural approach is to consider a class or a neighborhood of priors, and classes or collections of sampling distributions. This approach leads naturally to consideration of upper and lower probabilities or interval-valuedl probabilities.

We examine the various approaches to representation of imprecision in priors and sampling distributions. We realize that many useful classes, either for the priors or for the sampling distributions, are conveniently described in terms of 2-Choquet Capacities.

We prove the Bayes' Theorem (or Conditioning) for the 2-Choquet Capacity classes. Since the classes of imprecise probabilities described by the Dempster-Shafer Theory are ∞-Choquet Capacities (and therefore 2-Choquet Capacities) our result provides another proof of the inconsistency of the Dempster's rule.

We address the problem of combination of various sources of information and the requirements for a reasonable combination rule. Here, we also examine the

issues of independence of sources of information which is a crucial issue in combining various sources of information. We consider three methods to combine imprecise information. In method one, we utilizes the extreme-point representations of the imprecise priors and/or the sampling distributions to obtain the extreme-points of the class of posteriors. This method is usually computationally very demanding. Therefore, we propose a simple iterative procedure that allows direct computation of not only the posterior probabilities, but also many useful posterior related quantities such as the posterior mean, the predictive density that the next observation would lie in a given set, the posterior expected loss of a decision or an action, etc. Finally,, by considering the joint space of observations and parameters, we show that if this class of joint probabilities is a 2-Choquet capacity class, we can utilize our Bayes' Theorem found earlier to obtain the posterior probabilities. This last approach is computationally the most efficient method.

Finally, we address the problem of decision making with imprecise posteriors obtained from imprecise priors and sampling distributions. Even though, allowing imprecision is a natural approach for representation of lack of information, it sometimes leads to complications in decision making and even indeterminacies. We suggest a few ad-hoc rules to resolve the remaining indeterminacies. The ultimate solution in such cases is to simply gather more data.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Inference is the process of observing a sample or samples and drawing information about certain parameters of the underlying process. There are two distinct approaches to inference problems: one approach utilizes prior information, and the other is based solely on the observation samples. It is taken as given that prior information should be utilized whenever available. To this extent the Bayesian approach provides a sound and coherent way of *Combining* prior information, represented by prior probabilities and sampling distributions.

Decision-making problems are specific forms of inference problems. In decision making problems two other elements are added; namely a set of actions or decisions and a loss function indicating our subjective measure of losses between the decision or action made and the true value of the parameter under consideration. Even though proper selection of a loss function is very important, in this work we will not consider this problem and assume that an appropriate loss function is given.

In many real world problems, however, prior probabilities and/or sampling distributions may not be known precisely. For instance, in the early stages of outbreak of any new disease, with a small sample size, it is difficult if not impossible to obtain a precise model for the disease epistemology. Another example is the case of high sample dimensionality, where rarely if ever the

available data are adequate to define a precise model. See Hoffbech and Landgrebe (1993), Kim and Landgrebe (1991), Landgrebe (1993), Safavian and Landgrebe (1991), and Lee and Landgrebe (1993).

Our goal in this research is to consider situations where one can at best only describe a class or a neighborhood of priors, and classes or collections of sampling distributions. In such cases, we propose various methods to combine imprecise priors and sampling distributions and consider the problem of decision making with imprecise posteriors. The range of posterior quantities would be indicative of how *robust* the posterior quantity is with regards to variations of the priors and the sampling distributions.

1.2 Basic Approaches to Imprecision and Uncertainty

There are several basic approaches to handling imprecision. When the source of imprecision is linguistic in nature, *fuzzy set theory* has proved to be very useful. In contrast to regular set theory where an element either completely belongs to a given set or it has no membership in that set, in the fuzzy set theory one allows *partial* membership. Calculus of fuzzy set theory is developed by Zadeh (1965).

Another fundamentally different approach to allow for imprecision is to extend the concept of point-valued probability measures to set-valued (or Interval-valued) probability measures. This approach was first studied by Artstein (1972) and later further studied by Puri and Ralescu (1983) and Negoita and Ralescu (1987). Here, one assigns (compact, i.e., closed and bounded) intervals of values between 0 and 1 for events under consideration. Additivity of real-valued probability measures is preserved under this approach and is extended to "set Additivity". The Bayes Theorem is provided for the interval-valued probability measures [Negoita 1987]. The only major problem with this line of thinking is that, here probability measure of the sample space is not 1! Instead, all that is required is that probability measure of the sample space should be an interval including 1, i.e., $[a, 1]$, where $a < 1$. This is very counter-intuitive. As one would

expect, probability of a sample space should be ***exactly*** equal to 1; otherwise one could augment another outcome to the sample space and assign the remaining uncertainty of (1−a) to that outcome!

Also, one method to assign interval-valued probabilities to events is to use two sets of measures, P_1 and P_2, such that $P_1(A) < Pr(A) < P_2(A)$ for all events A. Where P_2 is any ordinary probability measure, and P_1 is any measure such that $P_1(A) < P_2(A)$ for all events A. Note that when $P_1(A) = P_2(A)$ for all events A., one gets the usual point-valued probability measures. This is a special case of "Intervals of Measures" considered by DeRobertis and Hartigan (1978). We will examine intervals of measures more carefully in the sequel.

The third approach is to consider "higher-order" probabilities. That is, suppose in a coin tossing experiment one does not feel comfortable with simply assigning probabilities, say, 0.5 for "heads" and 0.5 for "tails". It is suggested by Domotor (1981) and Kyburg (1988) that one can consider a second-order probability on the values of probabilities. For instance, one can assign a probability of 0.9 that probability of "heads" is going to be 0.5. There are several major problems with this approach. It is obvious that if one does not feel comfortable with assigning "first-order" probabilities, it is not obvious why one would feel comfortable in assigning the "second-order" probabilities. This leads to an endless argument: Assign "third-order" probabilities on the "second-order" probabilities, etc. Also, it has been shown by Kyburg (1988), that second-order probabilities have nothing to contribute to the analysis and representation of uncertainty. "The same ends can be achieved more simply, and without the introduction of novel machinery, by combing the "first" -and ".second" - order probabilities into a joint probability space, even if they are conceptually different kinds of probabilities."

Finally, the most natural and useful approach in modeling imprecision, the one that we will consider in detail in the sequel, is to consider classes or neighborhoods of probability measures. This approach is not new. It has been considered by Koopman (1940) and Boole (1884), among others.

Several different approaches could lead to consideration of classes of probability measures. The most obvious one is to start with a nominal model (or probability measure) and then consider a neighborhood around the nominal model described in terms of some appropriate metric. Or, for instance, using available data estimate a nominal model and then consider the confidence interval around the nominal model, etc. Or, suppose instead of having numerical values of probabilities, one only has some knowledge of *partial ordering* among the various probabilities. For example, suppose all we know is that disease 1 is *more likely* than disease 2 *and* disease 3 is 5 to 10 times more likely than disease 2, etc. This kind of available information leads to a class of probabilities all compatible with the above given information. Or, as Boole (1884) first observed, one may start with knowledge of probabilities of only *some of* the events. Then, again, one can construct a class of probability measures that will be compatible with the known probabilities. This approach was resurrected and extended by Dempster (Dempster 1968) and later by Shafer (1976).

It is obvious that consideration of classes of probability measures would directly lead to consideration of "upper" and "lower" probabilities. The difference between the "upper" and the "lower" probabilities indicate the robustness or sensitivity with respect to the class of probabilities considered.

1.3 Related Works

Even though robustness with respect to deviation in priors with *fixed* sampling distributions has been studied extensively in the literature (see Berger (1992) for a survey), very few studies has been performed to analyze the sensitivity of posteriors (and posterior related quantities) with respect to variations on *both* the priors and the sampling distributions.

Considering model robustness, Smith (1983) examines the parametric case, where a given model is "elaborated" or enlarged by considering a family of models parameterized with one (or more) new parameters. In our work, we

consider the non-parametric case and examine robustness with respect to both the priors and the models.

When both the parameter space and the measurement space are discrete, White (1986), considers classes of priors and sampling probability mass functions that are described in terms of *linear inequalities.* These classes are *convex polyhedrons.* He characterizes these convex polyhedrons via their *extreme points* and uses Bayes' Theorem to combine all the extreme points to obtain the extreme points of the posterior probabilities. We will examine this approach in detail in the sequel. This approach, in general, suffers from a high computational cost.

1.4 Statement of the Problem

We now formally state the problem that is solved in this thesis. First of all, we will implicitly assume the existence of densities and regular conditional probabilities as needed and we will ignore, as much as possible, all other measure-theoretic questions.

Let Θ represent the parameter space. We assume that $\Theta \in \Re$. Let $X \in \Re^d$ represent the measurement space, $f(x/\theta)$ denote the sampling density and $\Pi(\theta)$ denote the prior distribution on Θ. In order to avoid differentiating between "summation" and "integration", we will use the following notation:

$$\int_A \Pi(d\theta) = \begin{cases} \int_A \pi(\theta)d\theta & \text{if } \theta \text{ is continous;} \\ \sum_{\theta \in A} \pi(\theta) & \text{if } \theta \text{ is discrete,} \end{cases} \tag{1.4.1}$$

We assume that instead of having a precise prior probability distribution Π, we know that $\Pi \in \Gamma^{\Pi}$, and instead of knowing $f(x/\theta)$, $0 \in \Theta$, we know that for each 0, $f(x/\theta) \in \Gamma_\theta^f$, where Γ^{Π} is a class of admissible priors and Γ_θ^f are classes of

admissible sampling densities. Then, We like to find the following posterior related quantities:

$$\bar{\rho}(\phi,\Pi,f) \overset{\Delta}{=} \sup_{\Pi\in\Gamma_\Pi} \sup_{f(x/\theta)\in\Gamma_\theta'} \int_\Theta \phi(\theta) f(x/\theta)\,d\Pi(\theta) \Big/ \int_\Theta f(x/\theta)\,d\Pi(\theta) \tag{1.4.2}$$

$$\underline{\rho}(\phi,\Pi,f) \overset{\Delta}{=} \inf_{\Pi\in\Gamma_\Pi} \inf_{f(x/\theta)\in\Gamma_\theta'} \int_\Theta \phi(\theta) f(x/\theta)\,d\Pi(\theta) \Big/ \int_\Theta f(x/\theta)\,d\Pi(\theta) \tag{1.4.3}$$

Note that for the following choices of $\phi(\theta)$:

1) $\phi(\theta) = \theta$
2) $\phi(\theta) = I_B(\theta)$
3) $\phi(\theta) = \lambda(\theta,\delta(x))$

we have (1) the posterior mean, (2) the posterior probability of set B, and (3) the posterior expected loss of decision $\delta(x)$. The range $[\bar{\rho}-\underline{\rho}]$ indicates the degree of robustness or sensitivity of posterior quantity p with respect to the deviations in the priors and the sampling densities.

1.5 Thesis Organization

In Chapter 2, we examine several useful classes for priors and sampling distributions. These neighborhoods have very natural and useful interpretations. We examine both finite spaces and continuos spaces. We point out that most of these neighborhoods can be characterized in terms of 2-Choquet Capacities. Here, we formally introduce Choquet Capacities. An example of classes of uncertainty described by Choquet Capacities is the Dempster-Shafer (D-S) class. Therefore, we provide a brief exposure to the Dempster-Shafer Theory as we will be referring to this particular class frequently in the sequel.

Realizing the importance of 2-Choquet Capacity classes, we prove Bayes' Theorem (or conditional Choquet Capacities) for this class in Chapter 3. As mentioned earlier since D-S classes are ∞-Choquet Capacity classes (therefore, 2-Choquet Capacity classes), therefore our results apply there as well and furthermore provide another proof for the inconsistency of the Dempster's rule.

In Chapter **4,** we examine the Bayes' Theorem in statistical applications. First, we study the desired properties for any combination rule. Next, we investigate the issue of independence in combing sources of information and point out the potentials for assuming "too much" independence. Then, we examine properties of Bayes' rules and D-S combination rule in light of the enlisted properties and highlight the strength and weakness of each approach.

In Chapter 5, We provide (or introduce) three methods based on the Bayes' Theorem for combination of imprecise sources of information. In the first approach, we utilize the extreme point representation originally suggested by White (1986) and obtain the *posterior* extreme points from the extreme points of the priors and the sampling distributions. We look at the computational complexity of this approach and compare it to the computational complexity of D-S Theory. Even though the Bayesian approach has better computational complexity and does not suffer inconsistency criticisms of D-S Theory, here still computational complexity may be a problem. Thus we propose a second method that uses a linearization technique of Wasserman, Lavin and Wolpert (11393). This approach is iterative and converts a nonlinear optimization problem for finding $\bar{\rho}$ (or $\underline{p}$) into a sequence of simpler linear optimizations. We provide several examples here. As the third and final approach, we look at the product or the joint space of measurements and parameters, $\chi \times \Theta$. We note that if the class of joint distributions (or densities) is described in terms of a joint 2-Choquet Capacity, then we can utilize the Theorem of Chapter 3 and find the Posterior Choquet Capacities directly. This approach has the simplest computational complexity. We provide several examples.

In Chapter 6, we look at the problem of decision-making with imprecise probabilities. In general, even though representation of priors and sampling distributions in terms of classes of priors and sampling distributions is a natural way to indicate our available knowledge (or lack of it), this approach may sometimes lead to complications in decision-making, and even perhaps indeterminacies between certain actions or decisions. We provide a few *ad-hoc* suggestions to resolve possible cases of indeterminacies. The optimal solution, however, would be to simply acquire more data!

In Chapter 7, we provide our conclusions and directions for further research areas.

CHAPTER 2

REPRESENTATION OF IMPRECISE INFORMATION

2.1 Introduction

In this chapter, we will examine various approaches for representation of imprecision in priors and sampling distributions (also sometimes referred to as likelihoods, models, or conditional probabilities). We will look at the discrete case and the continuos case separately and motivate each method of representation with an example. We note that many of the useful and natural methods for describing imprecision can be characterized in terrns of 2-Choquet Capacities. We will formally introduce Choquet Capacities. As an important example of Choquet Capacities, we will consider the class described by the Dempster-Shafer Theory.

2.2 Discrete Case:

2.2.a Class of Imprecise Prior Probabilities:

Suppose Θ is the parameter space (e.g., space of all classes of interest in a classification problem, etc.) where $\Theta = \{\theta_1,...,\theta_M\}$. Without any loss of generality, in order to be able to provide a geometrical representation for demonstration, let us assume $M = 3$. Then the space of all possible priors is the probability simplex shown in Figure 1 below which, employing a system of triangular coordinates, can be displayed in 2-dimensions as in Figure 2.

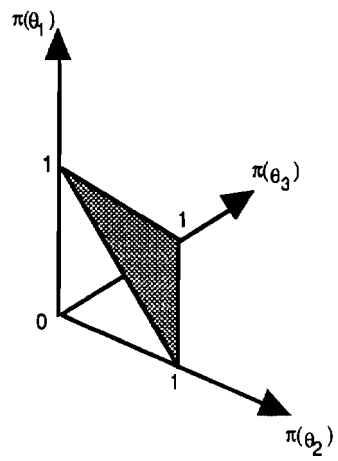

Figure 2.1 - Probability simplex in 3-d

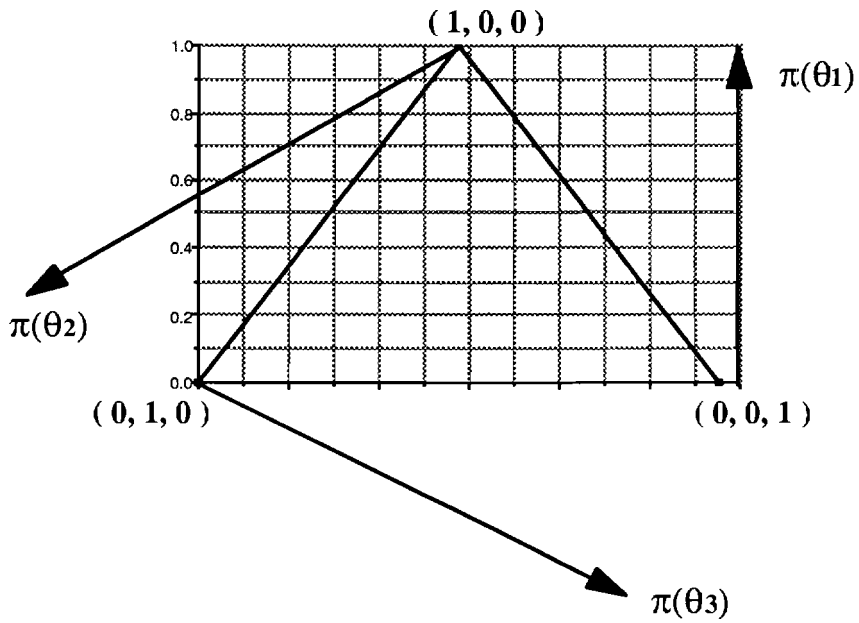

Figure 2.2 - Probability Simplex Using Triangular Coordinate System

Example 2.1: Suppose we do not have enough informatiori to construct a precise prior distribution for Θ, but we know that, for instance, the prior probability of θ_1 is at least 0.5 and θ_2 is more likely than *8,.* The class of priors corresponding to the above information is

$$\Gamma_1 = \left\{ \pi:\ \pi(\theta_1) \geq 0.5,\ \pi(\theta_2) \geq \pi(\theta_3),\ \sum_i \pi(\theta_i) = 1 \right\} \qquad (2.2.1)$$

which corresponds to the *convex* shaded area shown in Figure 2.3 below.

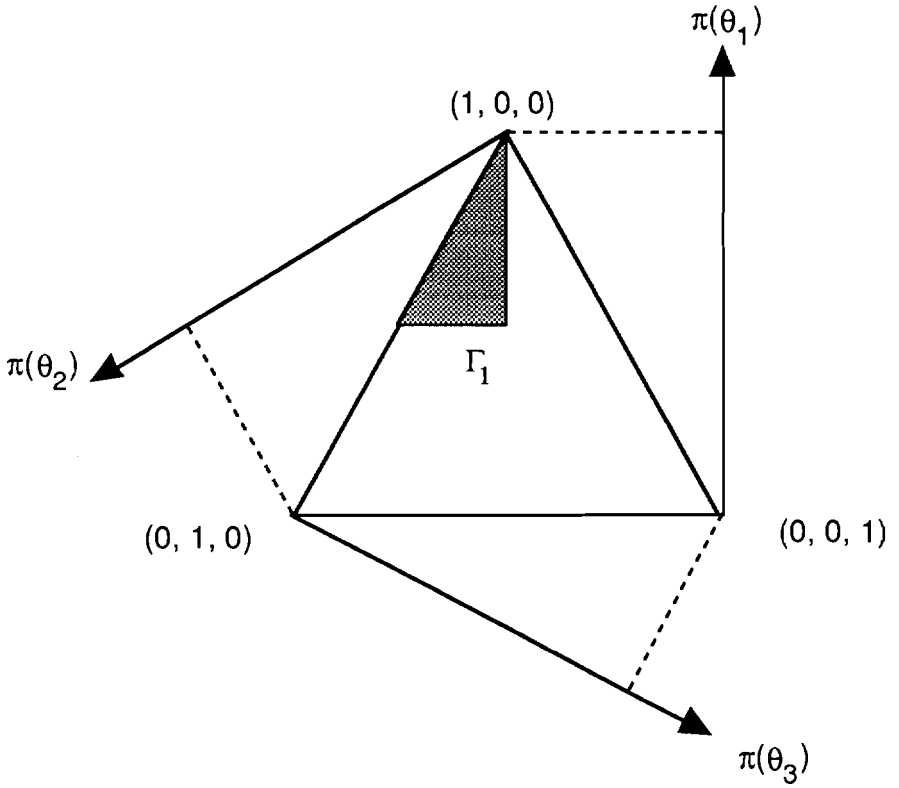

Figure 2.3 - Class of Priors for Example 2.1

Remarks:

1) The convex set Γ_1 can be completely specified in terms of its extreme points (or vertices).

2) The case of "total ignorance" or complete lack of knowledge would correspond to

$$\Gamma = \left\{ \pi:\ 0 \le \pi(\theta_i) \le 1,\ \sum_i \pi(\theta_i) = 1 \right\} \tag{2.2.2}$$

i.e., the entire simplex of probabilities. Even though the case of total ignorance occurs rarely in applications, the above representation is more natural and suitable than the conventional approach where one uses "non-informative" or a uniform distribution,

$$\pi(\theta_i) = \frac{1}{M} \quad ,\ i = 1, \ldots, M. \tag{2.2.3}$$

There are at least two problems with this latter representation. First, the uniform distribution does not exactly correspond to "total ignorance", as with the uniform distribution one expresses the knowledge that, for instance, θ_1 is as likely to occur as θ_2 where as in the case of total lack of knowledge this information is not available.

Second, for the case of a continuos and unbounded parameter space (e.g., $\Theta = \Re$), non-informative priors become improper priors, i.e.,

$$\int_{\Theta} d\Pi(\theta) = \infty \tag{2.2.4}$$

An interesting example indicating the inconsistencies that may arise using non-informative priors is provided by Shafer (1976, page. 24). See also Fishburn (1965) and Potter and Anderson (1980).

Another useful approach to represent imprecision is to specify lower and upper bounds for the prior probabilities $\pi(\theta_i)$. Where the lower and upper bounds indicate the minimum degree of belief or support, and the maximum degree of support for θ_i, respectively. That is, we consider the class

$$\Gamma_2 = \left\{ \pi:\ 0 \le l_i \le \pi(\theta_i) \le u_i \le 1,\ \sum_i \pi(\theta_i) = 1 \right\} \qquad (2.2.5)$$

Of course, it is possible that for some parameter(s), say θ_k, $l_k = u_k$, i.e., the prior probability $\pi(\theta_k)$ is known precisely .

It is straightforward to check that for Γ_2 to be non-empty, we need to have

R1) $$\sum_i l_i \le 1 \quad \text{and} \quad \sum_i u_i \ge 1 \qquad (2.2.6)$$

Consider the following example.

Example 2.2: Let $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and

$$0.0 \le \pi(\theta_1) \le 0.6$$
$$0.4 \le \pi(\theta_2) \le 0.5$$
$$0.2 \le \pi(\theta_3) \le 0.4$$

There are many probability distributions which obey these inequalities. For instance,

$$\pi(\theta_1) = 0.2\ ,\ \pi(\theta_2) = 0.4,\ \pi(\theta_3) = 0.4$$
or $$\pi(\theta_1) = 0.3\ ,\ \pi(\theta_2) = 0.4,\ \pi(\theta_3) = 0.3$$

Nevertheless the above interval representation is not satisfactory because $\pi(\theta_2) \ge 0.4$ and $\pi(\theta_3) \ge 0.2$ would in- ply that $\pi(\theta_1)$ can not be larger than 0.4., so the upper bound of 0.6 specified for $\pi(\theta_3)$ is unnecessarily too large. Similarly, since $\pi(\theta_2) \le 0.5$ and $\pi(\theta_3) \le 0.4$, this implies that $\pi(\theta_1)$ has to be larger than 0.1. Therefore the new lower and upper bounds for $\pi(\theta_1)$ are

$$0.1 \le \pi(\theta_1) \le 0.4$$

In other words, with the above original interval specifications, for

$$0.0 \le \pi(\theta_1) \le 0.1 \quad \text{or} \quad 0.4 \le \pi(\theta_1) \le 0.6$$

we can not find any probability distributions that satisfy the remaining constraints. That is, there are regions that are *infeasible.* We state this formally.

Definition 2.1: *A* non-empty class Γ of probability distributions is *feasible* if for each i, and every α_i with $l_i \le \alpha_i \le u_i$, there exists at least one probability distribution in Γ such that $\pi(\theta_i) = \alpha_i$.

The class of prior probabilities corresponding to example 2 is shown in Figure **2.4** below.

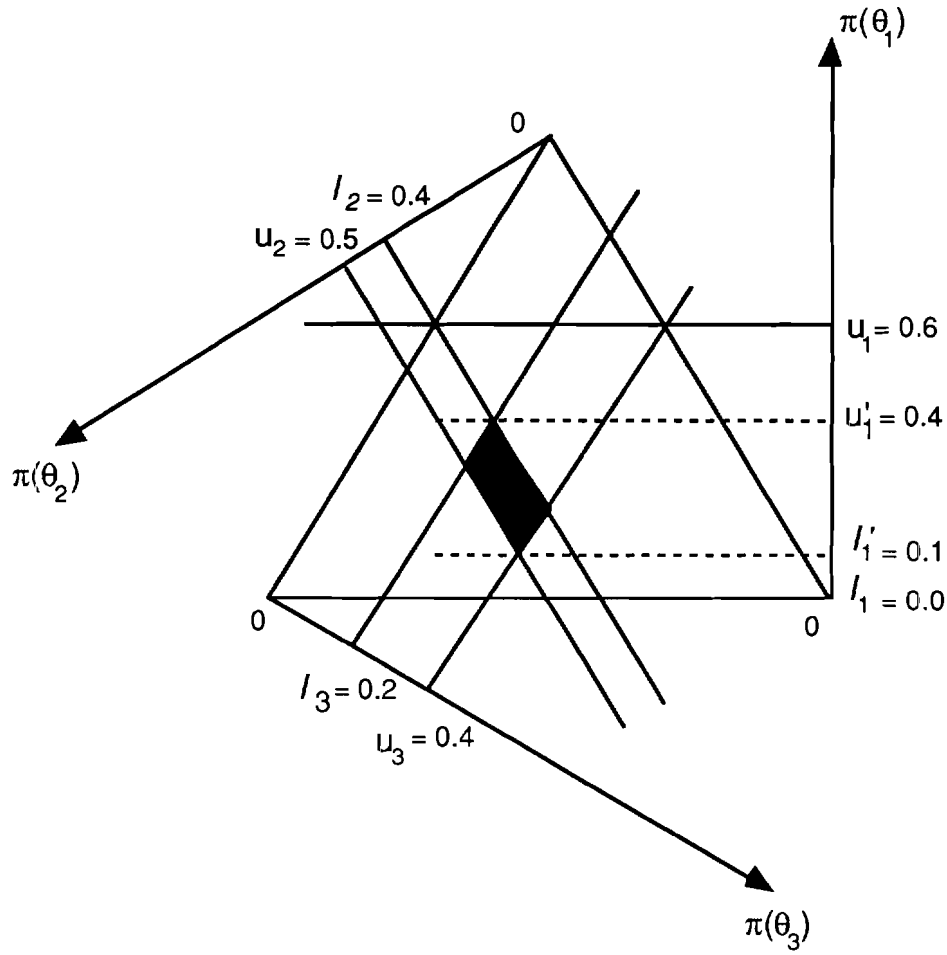

Figure 2.4 - Class of Priors for Example 2.2. Dark Area Is the Feasible Set. l_j' and u_j' Are the New Lower and Upper Bounds for $\pi(\theta_1)$.

It is easy to verify that a given class Γ is feasible if and only if the lower and upper bounds satisfy the following requirements:

R2)
$$\begin{cases} u_j \le 1 - \sum_{\substack{i=1 \\ i \neq j}}^{M} l_i \\ l_j \ge 1 - \sum_{\substack{i=1 \\ i \neq j}}^{M} u_i \end{cases} \quad \text{for j =1, ..., M.} \qquad (2.2.7)$$

Therefore given an arbitrary set of upper and lower bound specifications, we need to first check the requirement R1) to make sure that the class Γ is non-empty and check the requirement R2) to verify that the bounds specified are not too large. In case the bounds are too large, we can refine them using the following result.

Lemma 2.1: Given a non-empty *infeasible* class Γ, the new lower and upper bounds for the *feasible* class are given by

$$l_j' = \max\{l_j, 1 - \sum_{\substack{i=1 \\ i \neq j}}^{M} u_i\} \tag{2.2.8}$$

$$u_j' = \min\{u_j, 1 - \sum_{\substack{i=1 \\ i \neq j}}^{M} l_i\} \tag{2.2.9}$$

We will refer to the feasible class Γ_2 as the *discrete band model.* Band models in general play an important role in our studies.

The following classes can be used for *both* finite (discrete) spaces and continuos spaces.

Suppose we have elicited a nominal prior, say $\pi_o(\theta)$, but we do not feel 100% certain about it. Suppose, however, we feel (1-ε)% comfortable with this nominal model, where $0 \leq \varepsilon \leq 1$. This type of information can be conveniently described as:

$$\Gamma_3 = \left\{ \pi:\ \pi = (1-\varepsilon)\,\pi_o + \varepsilon\, q \right\} \tag{2.2.10}$$

where q can be any arbitrary distribution (referred to as contamination). This class is known as the *ε-contamination* class and was introduced by Huber (1973) and Berger and Berliner (1986).

In the *ε-contamination* class one needs knowledge of a nominal model to work with. In the absence of a nominal model, we can consider the following class

$$\Gamma_4 = \left\{ \pi:\ l(\theta) \le \pi(\theta) \le u(\theta);\ \int_{\Theta} \pi(\theta) d\theta = 1 \right\} \qquad (2.2.11)$$

where $\int_{\Theta} l(\theta) d\theta \le 1 \le \int_{\Theta} u(\theta) d\theta$, with the equalities corresponding to the trivial case where the prior is known precisely. This is known as the density *bounded* model and was introduced by Kassam (1981) and Lavine (1991). Mote that in this class, one allows the prior to have *any shape* as long as it is bounded from below and above with $l(\theta)$ and $u(\theta)$, respectively.

Even though to use a band model one does not need a nominal model, band models can also be used in situations where one estimates a model from the available data and then considers a pair of confidence limits around this model. Furthermore, another specialization of the band model is obtained by taking $l(\theta) \equiv 0$, in which case the class is completely characterized in terms of the upper bound $u(\theta)$ only.

In the density bounded model, one only considers valid densities that are bounded by $l(\theta)$ and $u(\theta)$. A simple but useful generalization of density bounded models can be obtained as

$$\Gamma_5 = \left\{ \pi = \frac{f}{\int_{\Theta} f}:\ \ l(\theta) \le f(\theta) \le u(\theta) \right\} \qquad (2.2.12)$$

This is known as the *density ratio* class or the *band model'* class and was introduced by DeRoberis and Hartigan **(1981).** Here, one considers *all* the functions $f(\theta)$ that are bounded by $l(\theta)$ and $u(\theta)$, and then normalizes them to get valid densities.

Finally, we consider the following class known as the *total variation* class. This is the class of all prior measures that are at most ε-away from the nominal measure Π , where distance is measured by the metric *d:*

$$\Gamma_6 = \left\{ \Pi:\ d\left(\Pi(A),\ \Pi_o(A) \right) \leq \varepsilon \right\} \tag{2.2.13}$$

where *d* could be either the total variation, Prohorov, Kolmogorov, or Levy distance, and ε is a fixed constant, $0 < \varepsilon < 1$.

2.2.b Classes of Imprecise Sampling Distributions

All the classes introduced above can also be adapted to represent imprecision about the sampling distributions. For instance, suppose we! have nominal sampling densities $f_\theta(x/\theta),\ 8 \in \Theta$ which might have been estimated from an available training sample. Then, we can consider, for example, the *ε-contamination* classes

$$\Gamma_\theta^f = \left\{ f(x/\theta):\ f(x/\theta) = (1-\varepsilon_\theta) f_O(x/\theta) + \varepsilon_\theta\ q(x/\theta) \right\} \tag{2.2.14}$$

where $(1-\varepsilon_\theta)$ reflects our confidence in the nominal model $f_o(x/\theta)$, and for the sake of generality, we have allowed the different sampling distributions to have different degrees of contamination depending on θ. We could have also considered the *density bounded* model, i.e.,

$$\Gamma_{\theta}^{f} = \left\{ l(x/\theta) \le f(x/\theta) \le u(x/\theta);\ \int_{X} f(x/\theta)dx = 1 \right\} \qquad (2.2.15)$$

where $l(x/\theta)$ and $u(x/\theta) \ge 0$ and $\int l(x/\theta)dx \le 1 \le \int u(x/\theta)dx$, etc.

2.3 Choquet Capacities

Except for the class Γ_1 which is a general *convex* set and does not necessarily have any other structure, the remaining classes have richer structure and can all be characterized in terms of Choquet Capacities. Next, we formally define Choquet Capacities.

Let Ω be a sample space and a be a Borel field (or o-algebra) on Ω. If Ω is finite, then we can take a to be the power set. Then any set function defined on a that satisfies the following properties *p1) - p4)* is called a Choquet Capacity (Choquet (1953) and Huber (1973)):

$$\begin{aligned}
&p1) \quad v(\phi) = 0, \qquad v(\Omega) = 1 \\
&p2) \quad A \subseteq B \quad \Rightarrow \quad v(A) \le v(B) \\
&p3) \quad A_n \uparrow A \quad \Rightarrow \quad v(A_n) \uparrow v(A) \\
&p4) \quad F_n \downarrow F\ ,\ F_n closed \quad \Rightarrow \quad v(F_n) \downarrow v(F)
\end{aligned}$$

If it also satisfies the *sub-Additivity* property P5) below,

$$p5) \quad v(A \cup B) + v(A \cap B) \le v(A) + v(B)$$

then it is called an alternating of order 2, or for short, *2-alternating* capacity. More generally, a Choquet capacity that satisfies

$$v(\bigcup_{i=1}^{n}) \leq \sum_{\substack{I \neq \phi, \\ I \subset \{1,\ldots,k\}}} (-1)^{|I|+1} v(\bigcap_{i \in I} A_i) \qquad (2.3.1)$$

i.e.,

$$v(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} v(A_i) - \sum_{i<j} \sum v(A_i \cap A_j) + \sum_{i<j<K} \sum \sum v(A_i \cap A_j \cap A_k) + \ldots$$
$$+ (-1)^{|I|+1} v(A_1 \cap \ldots \cap A_n)$$

is called an *n-alternating* capacity. If it satisfies the above relationship for *any* n, then it is called an ∞ - *alternating* capacity.

Note that property p1) is just the boundary condition, *p2)* is the monotonicity, *p3)* and *p4)* are continuity conditions from below and above for arbitrary increasing sequences of events and decreasing sequences of events that are closed sets, respectively. And *p5)*is a weak form of Additivity. Similarly, a set function u that satisfies p1') – *p4')*

$$\begin{aligned}
&p1') \quad u(\phi) = 0, && u(\Omega) = 1 \\
&p2') \quad A \subseteq B && \Rightarrow \quad u(A) \leq u(B) \\
&p3') \quad A_n \downarrow A && \Rightarrow \quad u(A_n) \downarrow u(A) \\
&p4') \quad G_n \uparrow G \text{ , } G_n open && \Rightarrow \quad u(G_n) \uparrow u(G)
\end{aligned}$$

and the *super-Additivity* property *p5')*

$$p5') \quad u(A \cup B) + u(A \cap B) \geq u(A) + u(B)$$

is called a monotone of order 2, or for short, *2-monotone* capacity. Similarly, if a monotone Choquet capacity satisfies

$$u(\bigcup_{i=1}^{n} A_i) \geq \sum_{\substack{I \neq \phi, \\ I \subset \{1,\ldots,k\}}} (-1)^{|I|+1} u(\bigcap_{i \in I} A_i) \qquad (2.3.2)$$

i.e.,

$$u(\bigcup_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} u(A_i) - \sum_{i<j} \sum u(A_i \cap A_j) + \sum_{i<j<K} \sum \sum u(A_i \cap A_j \cap A_k) + \ldots$$
$$+ (-1)^{|I|+1} u(A_1 \cap \ldots \cap A_n)$$

is called an *n-monotone* capacity. If it satisfies the above relationship for *any* n, then it is called an ∞ - *monotone* capacity.

Remarks:

1) Alternating and monotone capacities v and u, satisfy

$$v(A) + u(A^c) = 1 \qquad (2.3.3)$$

and are said to be *conjugates.* Therefore, it suffices to consider only one of these two functions.

2) It is known that if u (v) is monotone (alternating) of order *n*, then it is also monotone (alternating) of order *k* for any integer $2 \leq k \leq n$.

3) Probability measures are special types of capacities: they are *both* ∞ - *monotone* and ∞ - *alternating* capacities.

Next, we will provide some motivation for the above Choquet capacity definitions. Given a measurable space (Ω, A), let M denote the set of all probability measures on Ω, and, let P be a non-empty subset of M; $\mathcal{P} \subset \mathcal{M}$. Then, one may define the following *lower* and *upper probabilities* induced by P:

$$v(A) = \sup_{P \in \mathcal{P}} P(A), \qquad A \in \mathcal{A}, \tag{2.3.4a}$$

and

$$u(A) = \inf_{P \in \mathcal{P}} P(A), \qquad A \in \mathcal{A}. \tag{2.3.4b}$$

It is easy to see that

$$v(A) + u(A^c) = 1. \tag{2.3.5}$$

That is, these two set functions are conjugate pairs. The set functions u and v are called a "lower envelope" for $\mathcal{P}$ and an "upper envelope" for $\mathcal{P}$, respectively. Note that when the true probability distribution P is unknown, and it is only believed that $P \in \mathcal{P}$, then u and v provide us with a lower and upper bound for the actual value of the unknown probabilities. The interval $[u \ v]$ is called an "interval-valued probability"; Kim (1990).

More importantly, Huber and Strassen (1973) have shown that if $\mathcal{P}$ is weakly compact, then the set functions u and v are capacities (though not necessarily of any order).

Conversely, one may start with an arbitrary pair of conjugate capacities u and v, and define the sets $\mathcal{P}_v$, $\mathcal{P}_u$, and $\mathcal{P}_{uv}$ by:

$$\mathcal{P}_v = \{P \in \mathcal{M} / P(A) \leq v(A), A \in \mathcal{A}\} \tag{2.3.6a}$$

and $$\mathcal{P}_u = \{P \in \mathcal{M} / u(A) \leq P(A), A \in \mathcal{A}\} \tag{2.3.6b}$$

$$\mathcal{P}'_{uv} = \{P \in \mathcal{M} / u(A) \leq P(A) \leq v(A), A \in \mathcal{A}\} \qquad (2.3.6c)$$

It is known that $\mathcal{P}_v = \mathcal{P}_u = \mathcal{P}'_{uv}$. So we can only consider one of them; say $\mathcal{P}_v$. The set $\mathcal{P}_v$ is the set of all probability measures *dominated* by v. It is pointed out by Huber (1981) that in general (2.3.4) followed by (2.3.6) does not restore P, that is in general $\mathcal{P}_v \neq \mathsf{T}$. If $\mathcal{P}_v = \mathcal{P}$, then we call the set function v and the set of probabilities P *representable.*

Huber and Strassen (1973) have shown that only when v (or *u)* are alternating (or monotone) of order 2 or higher that we have this useful property. This further emphasizes the importance of 2-capacities, that is, alternating or monotone capacities of order 2. As we will see next, fortunately, almost all of the classes for describing imprecision introduced earlier can be represented with an appropriate 2-capacity. Our discussion will be around the classes of imprecise priors, but as we mentioned earlier, with only a slight change of notation, the same argument will hold for the classes of imprecise sampling distributions.

In particular, the ε-contamination, the band model or the generalized band model, the density-ratio, the total variation, the Prohorov, etc. classes can be all represented by some appropriate 2-capacities.

The ε-contamination class:

$$\Gamma_{\varepsilon c} = \left\{ \Pi:\ \Pi(A) = (1-\varepsilon)\Pi_o(A) + \varepsilon Q(A) \right\} , \qquad (2.3.7)$$

where $\Pi_o(A)$ is a nominal prior measure, $0 < \varepsilon < 1$, and Q is *any* arbitrary (contaminating) measure, can be represented by

$$\Gamma_{\varepsilon c} = \{ \Pi:\ \Pi(A) \leq v(A) \} \qquad (2.3.8)$$

where

$$v(A) = (1-\varepsilon)\Pi_o(A) + \varepsilon \tag{2.3.9}$$

is a 2-alternating capacity.

The density bounded class:

$$'\textbf{density bounded} = \{\ \Pi:\ L(A) \le \Pi(A) \le U(A)\ ;\ \Pi(\Theta) = 1\ \} \tag{2.3.10}$$

where L and U are lower and upper measures (with densities l and u with respect to an appropriate measure and $L(\Theta) \le 1 \le U(\Theta) < \infty$). This class can be represented by

$$'\textbf{density bounded} = \{\ \Pi:\ \Pi(A) \le v(A)\ \} \tag{2.3.11}$$

where

$$v(A) = min\left\{\ U(A),\ 1 - L(A^C)\ \right\} \tag{2.3.12}$$

is a 2-alternating capacity. A^C denotes the complement of the set A.

The total variation class:

$$\Gamma t\text{-}v = \left\{\ \Pi:\ \left|\Pi(A) - \Pi_O(A)\right| \le \varepsilon\ \right\} \tag{2.3.13}$$

where $\Pi_o(A)$ is a nominal prior measure, and $0 < \varepsilon < 1$ can be represented by

$$\Gamma_{t\text{-}v} = \{\ \Pi:\ \Pi(A) \le v(A)\ \} \tag{2.3.14}$$

and

$$v(A) = min\left\{\ \Pi_O(A) + \varepsilon\ ,\ 1\ \right\}. \tag{2.3.15}$$

is a 2-alternating capacity, etc.

Next, we consider an important family of ∞- capacities arising from Dempster-Shafer (D-S) theory. We start with a brief introduction to D-S theory.

2.4 Dempster-Shafer Theory

The basic idea can become clear with the following simple (desk) example. Suppose there is a desk with two drawers on the right side: the right top drawer (RT) and the right bottom drawer (RB) . There are three drawers on the left side: the left top drawer (LT), the left middle drawer (LM), and the left bottom drawer (LIB). So, the sample space is Ω = { RT, RB, LT, LM, LB}.

Suppose a file is placed, at random, in one of the drawers. Further suppose that the available information (or evidence in the D-S language) is given as

$$prob(\text{ file is in any of the left side drawers}) = m\ (\text{LT} \cup \text{LM} \cup \text{LB}) = 0.5$$
$$\text{prob}(\text{ file is in the RT drawer }) = m(\text{ RT }) = 0.2$$

and there is no more information.

Note that the total evidence, $m(\text{LT} \cup \text{LM} \cup \text{LB}) + m\ (\text{RT}) = 0.7 < 1$. Shafer calls the difference (1– 0.7 =0.3), the global ignorance . The global ignorance can be assigned to any of the drawers (sets), and yet to none in particular. Then given the above scenario, one would like to answer questions like: what is the probability that the file is in the (LM) drawer, etc. Obviously, the answer to these questions can not be given by single numbers. George Boole [4] was the first to realize this point and he suggested the idea of inner and outer measures, p. and p', such that probability of any event, p, is bounded by p_* and p^* as

$$p_* \le p \le p^*$$

Shafer calls $m(.)$ the *basic probability assignments* or (bpa)'s. $m(A)$ represents the measure of belief that is committed *exactly* to set A and not to any of its

proper subsets. Note that if $m(.)$ can be specified for every singleton, then bpa reduces to the usual probability mass function. bpa is formally defined as:

DEFINITION 2.2: [Shafer (1976)]

A function m: $2^{\Omega} \rightarrow [0,1]$, where 2^{Ω} is the power set of Ω, is called a basic probability assignment (bpa) whenever

$$\begin{aligned} &(1) \quad m(\phi) = 0 \\ &\text{and} \\ &(2) \quad \sum_{A \subset \Omega} m(A) = 1 \end{aligned} \tag{2.4.1}$$

Note that

i) It is not required that $m(\Omega) = 1$;
ii) It is not required that $m(A) \leq m(B)$ when $A \subseteq B$;
iii) There is no obvious relationship between $m(A)$ and $m(A^c)$.

Recall that $m(A)$ reflects the measure of belief that is committed exactly to A, not the *total* belief that is committed to A. To obtain the total belief committed to A, Shafer argues, that one must add to $m(A)$, the bpa of all the proper subsets B of A. He calls this belief function or Bel for short. That is

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2.4.2}$$

Dampster in his original work called these Bel functions, lower probabilities. More formally, a function Bel: $2^{\Omega} \rightarrow [0,1]$ is called a belief function if it is given by (2.4.2), for some bpa m: $2^{\Omega} \rightarrow [0,1]$. For our earlier "desk" example :

Bel ('file is in (ML) drawer) = 0.

Bel('file is in (RT) drawer) = 0.2

It is important to note that

$$Bel(A) + Bel(A^c) \leq 1 \tag{2.4.3}$$

To see the implication of this relationship, suppose that there is no evidence at all to support A, or $Bel(A) = 0$. Then, (2.4.3) says that, in D-S theory, it is not automatically implied that $Bel(A^c) = 1$; i.e., lack of belief in something does not necessitate its compliment.

Furthermore, the bpa that produces a given belief function can be *uniquely* recovered from the belief function. This inverse relation is called *mobius* inverse. For any belief function Bel, a dual function called the plausibility function (or "Pl" for short) is defined as

$$Pl(A) = 1 - Bel(A^c) \tag{2.4.4}$$

In terms of bpa m, plausibility could be written as

$$Pl(A) = \sum_{B \cap A \neq \phi} m(B) \tag{2.4.5}$$

Dempster called these Pl's, upper probabilities. Note

$$Pl(A) + Pl(A^c) \geq 1 \tag{2.4.6}$$

and

$$Pl(A) \geq Bel(A) \tag{2.4.7}$$

From our earlier "desk" example:

PI (file is in (ML) drawer) = 0.3
PI (file is in (RT) drawer) = 0.5 .

To make the idea of "*Bel*" and "Pi" clearer, let us consider the following example. Suppose we are given: $m(B_1) = 0.3$, $m(B_2) = 0.4$, and $m(B_3) = 0.1$, thus $m(\Omega) = 0.2$, and want to find the lower and upper probability (or *Bel* and *Pl*) of a set A given in the following diagram.

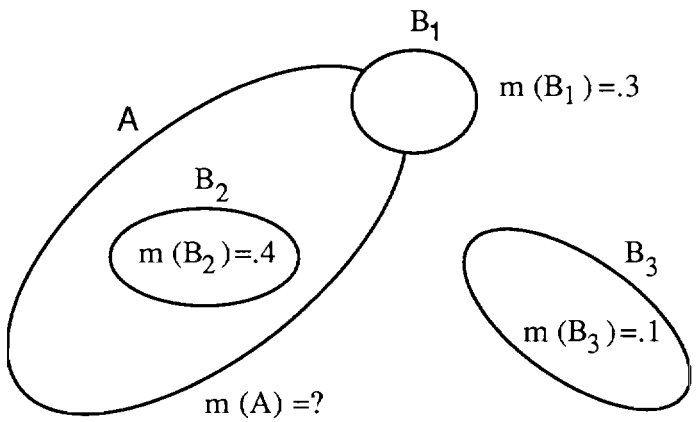

Figure 2.5 - Some Arbitrary Sets with Their Associated BPA Numbers

$$Bel(A) = \sum_{B_i \subseteq A} m(B_i) = m(B_2) = 0.4$$

$$Pl(A) = \sum_{B_i \cap A \neq \phi} m(B_i) = m(B_1) + m(B_2) + m(\Omega)$$

$$= 0.3 + 0.4 + 0.2 = 0.9$$

Shafer, further argues that the class of belief functions can be characterized without reference to any basic probability assignment function. That is:

THEOREM 2.1: [Shafer (1976)]

A function $Bel: 2^{\Omega} \rightarrow [0,1]$ is *a belief function if and only if it satisfies the following:*

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

(2) $Bel(\Omega) = 1$

(3) *For every positive integer* n *and every collection* $A_1, \ldots, A_n$ *of subsets of* Ω

$$Bel(A_1 \cup \ldots \cup A_n) \geq \sum_i Bel(A_i) - \sum_{i<j} Bel(A_i \cap A_j) + \ldots + (-1)^{n+1} Bel(A_1 \cap \ldots \cap A_n). \quad (2.4.8)$$

Remark: Note that Bel functions are ∞ - monotone capacities .

As mentioned earlier, there is a one-to-one correspondence between basic probability assignments (bpa) and Bel functions; i.e., given a bpa one can construct the corresponding Bel function and conversely given a Bel function one can obtain the corresponding bpa. This relationship is called mobius inverse. More precisely,

THEOREM 2.2: [Shafer (1976)]

Suppose $Bel: 2^{\Omega} \rightarrow [0,1]$ *is the* Bel *function given by its bpa* $m: 2^{\Omega} \rightarrow [0,1]$. *Then*

$$m(A) = \sum_{B \subset A} (-1)^{|A-B|} Bel(B) \quad (2.4.9)$$

for all $A \subset \Omega$.

Similarly, one can define plausibility functions as:

THEOREM 2.3: [Shafer (1976)]

A function $Pl: 2^{\Omega} \rightarrow [0,1]$ *is a plausibility function if and only if it satisfies the following conditions:*

(1) $Pl(\phi) = 0.$

(2) $Pl(\Omega) = 1.$

(3) *For every positive integer* n *and every collection* $A_1, \ldots, A_n$ *of subsets of* Ω

$$Pl(A_1 \cup \ldots \cup A_n) \leq \sum_i Pl(A_i) - \sum_{i<j} Pl(A_i \cup A_j) + \ldots + (-1)^{n+1} Pl(A_1 \cup \ldots \cup A_n). \qquad (2.4.10)$$

Remarks;

1) Note that Pl functions are ∞- alternating capacities.

2) When $Bel(A \cup B) = Bel(A) + Bel(B)$, $A \cap B = \phi$ belief function becomes the usual classical probability measures. Furthermore, one can show that (Klir [23]) a belief function, Bel, on a finite power set 2^{Ω} is a probability measure if and only if its basic probability assignment, m, is given by $m(\{\omega\}) = Bel(\{\omega\})$ and $m(\{A\}) = 0$ for all subsets of Ω that are not singletons.

3) A Bel function that satisfies $Bel(A) = 0$ for every proper subset A of Ω is called *avacuous* belief function. In terms of basic probability assignments, this means $m(\Omega) = 1$ and $m(A) = 0$ for every proper subset A of Ω . Furthermore plausibility of every such A is one. That is

$$Bel(A) = 0 \leq pr(A) \leq Pl(A) = 1 \qquad \forall A \subset \Omega.$$

Again, the major conclusion of this section is that Bel functions and Pl functions that are major components of D-S theory are ∞- monotone and ∞- alternating capacities, respectively.

In the next Chapter, we will drive the Bayes' Theorem for Capacities and compare the results with the corresponding rule given by D-S theory.

CHAPTER 3

BAYES' THEOREM FOR CAPACITIES

3.1 Bayes' Theorem in Probability

Consider a measurable space $(\Omega, \mathcal{A})$ along with a probability measure $pr(.)$ defined on a. Then, Bayes' theorem (or conditioning) in probability, in its simplest form, states that given the information that event B,$B \in \mathcal{A}$, has occurred, we need to revise our original belief function (expressed by $pr(.)$) as

$$pr(A/B) = \frac{pr(A \cap B)}{pr(B)} = \frac{pr(B/A)pr(A)}{pr(B)} \qquad (3.1.1)$$

provided $pr(B) > 0$. Where now the new sample space is B and A belongs to the appropriate o-algebra restricted to B. Here $pr(A')$ represents our knowledge about A' *before* observing B, $pr(B/A')$captures the relationship between A' and B, and $p(A'/B)$represents our new belief in A' *after* observing B, also referred to as the *posterior* belief function. Our goal next is to drive a relationship similar to the above but for Capacities.

3.2 Bayes' Theorem for Capacities

Suppose we have the same measurable space $(\Omega, \mathcal{A})$ but we are unable or unwilling to represent our beliefs via a precise probability measure. Instead we have chosen to consider a family or a neighborhood of probability measures

such as the ε - contamination family or the band model described in the Chapter 2. Recall that these neighborhoods could be expressed as

$$\mathcal{P} = \{pr: pr(A) \leq v(A)\} \ , \ A \in \mathcal{A} \tag{3.2.1}$$

and $v(.)$ is the 2-alternating capacity corresponding to (or representing) the neighborhood $\mathcal{P}$.

Now suppose we have observed B and wish to revise our beliefs in light of this new piece of information. Let $\mathcal{P}_{(.|B)}$ represent the family of revised or posterior measures. A simple but naive approach to obtain $\mathcal{P}_{(.|B)}$ would be to revise every probability measure $pr \in \mathcal{P}$. Of course, in most cases, this would be computationally prohibitive. Instead, we focus on the 2-alternating capacity $v(.)$. We first drive the conditional capacity $v(./B)$. Then, we consider the set of probability measures dominated by this new conditional capacity; i.e.,

$$\mathcal{P}_{v(./B)} = \{p: p(A') \leq v(A'/B))\} \tag{3.2.2}$$

In general, $\mathcal{P}_{v(./B)}$ would be somewhat larger than $\mathcal{P}_{(.|B)}$, providing a somewhat conservative estimate of the actual $\mathcal{P}_{(.|B)}$, but would have the advantage of providing a closed form solution.

3.3 Conditional Capacities

Before we prove the conditional capacity theorem, we need the following lemma due to Huber (1981, page 273):

Lemma:

Let $\mathcal{P}$ be a family of probability measures majorized with a 2-alternating capacity v; i.e.,

$$\mathcal{P} = \{pr: pr(A) \le v(A)\} \tag{3.3.1}$$

Then for any monotone sequence $A_1 \subset A_2 \subset \ldots \subset A_n$ *belonging to* $\mathcal{A}$, *it is possible to find a probability measure* $pr^* \in \mathcal{P}$, *such that simultaneously for all* i, $i = 1,2,\ldots,n$

$$pr^*(A_i) = v(A_i). \tag{3.3.2}$$

We also need the following facts:

1) Any set B can be decomposed into two disjoint sets: $B = (B \cap A) \cup (B \cap A')$.

2) $(B \cap A^c)^c = B^c \cup A$.

3) $\inf_{pr \in \mathcal{P}} pr(A) = \inf_{pr \in \mathcal{P}} (1 - pr(A^c)) = 1 - \sup_{pr \in \mathcal{P}} pr(A^c)$.

Now we are ready to state our theorem:

Theorem:

Let $\mathcal{P}$ *be a family of probability measures majorized with a 2-alternating capacity* v; *i.e.,*

$$\mathcal{P} = \{pr: pr(A) \le v(A)\} \tag{3.3.3}$$

and let

$$v(A/B) = \sup_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B)} \tag{3.3.4}$$

and

$$u(A/B) = \inf_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B)} \tag{3.3.5}$$

denote the upper and lower conditional probabilities over the family $\mathcal{P}$, respectively. Then, $v(A/B)$ and $u(A/B)$ can be expressed in terms of the original unconditional 2-alternating capacity v and its conjugate u as

$$v(A/B) = \frac{v(A \cap B)}{v(B \cap A) + u(B \cap A')}$$

and

$$u(A/B) = \frac{u(A \cap B)}{u(B \cap A) + v(B \cap A')}$$

Proof: See Appendix A.1.

Remarks:

1) It is easy to check to see that the conditional capacities $v(A/B)$ and $u(A/B)$ satisfy the properties $p1)$ - $p4)$ and $p1')$ - $p4')$, respectively. Therefore, they are both capacities.

2) For the finite sample space case, recently Sundberg and Wagner (1994a), (1994b) using a completely different line of reasoning have shown that the conditional capacities are also 2-alternating and 2-monotone capacities, respectively.

3) We *conjecture* that even for general sample spaces, the resulting conditional capacities will remain 2-alternating and 2-monotone capacities, respectively. But, we have not been able to show this yet.

Next, we compare our results with the conditional belief and plausibility functions given by D-S theory.

3.4 Conditioning in Dempster-Shafer Theory

Recall that lower and upper probabilities in the D-S framework are called the Belief function, Bel, and the Plausibility function, Pl, respectively; where Bel is an ∞-monotone capacity and Pl is an ∞-alternating capacity. Then, given an event B, the *conditional Bel* and *conditional Pl* derived directly from *Dempster's rule of combination of evidences* (Shafer 1976, page 66-67) are

$$Bel(A|B) = \frac{Bel(A \cup B^c) - Bel(B^c)}{1 - Bel(B^c)} \tag{3.4.1}$$

and

$$Pl(A|B) = \frac{Pl(A \cap B)}{Pl(B)} \tag{3.4.2}$$

Since Bel and Pl are conjugates, we need to examine only one of the above conditional quantities. Pl has a simpler form, so let us examine it. It is obvious that in general

$$Pl(A|B) = \sup_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B)} \neq \frac{\sup_{p \in \mathcal{P}} pr(A \cap B)}{\sup_{p \in \mathcal{P}} pr(B)} = \frac{Pl(A \cap B)}{Pl(B)} \tag{3.4.3}$$

A similar argument applies for the conditional Bel. This discrepancy provides another proof of inconsistency of Dempster's rule of combination.

Remark:

Note that the bound provided by D-S conditional Bel and Pl is in general tighter than the bound given by the conditional capacities; i.e.,

$$u(A|B) \leq Bel(A|B) \leq Pl(A|B) \leq v(A|B). \tag{3.4.4}$$

That is, the conditional capacities provide a more conservative estimate of the true conditional probabilities. See also Kyberg (1987).

CHAPTER 4

COMBINATION OF IMPRECISE SAMPLING DENSITIES AND IMPRECISE PRIORS

4.1 Introduction

The main objective of this chapter is to address the problem of combination of imprecise sampling distributions $\{P(x/\theta), \theta \in \Theta\} \in \Gamma_\theta^p$ with imprecise priors $\pi(\theta) \in \Gamma_{prior}$. To avoid unnecessary measure-theoretic issues, we will assume that all distributions under consideration have their corresponding densities (with respect to some appropriate measure, e.g., Lebesgue measure), thus we will consider imprecise sampling ***densities*** (also known as conditional densities, models, and likelihood functions) $\{f(x/\theta), \theta \in \Theta\} \in \Gamma_\theta^f$. We start by considering the conventional Bayesian approach and how it combines information provided by models and priors. Then we examine the Bayesian solution for combination of several sources of information where each source, S_i, is described in terms of a different (possibly imprecise) families of sampling densities, $\{f(x/\theta; S_i), \theta \in \Theta, i = 1,...,L\}$. Here, we need to closely examine the roe of assumption of "independence" and consequences of making "too much" independence assumption. We also briefly look at the notion of independence in the context of D-S theory and its consequences. See also Kim (1990), Benediktson, Swain and Ersoy (1989), and Lee, Richards and Swain (1987).

4.2 Independence and Combination of Sources of Informatiori

The major goals of this section are the following. First, we want to investigatethe desired properties that *any* rule for combination of information from various sources should have. Then, we will examine how Bayes' rule combines information under various types of independence assumptions and the resulting properties of these rule. Finally, we will briefly study how D-S theory combines information, some of its properties, the type of independence assumptions made there and their consequences. To keep our discussion general, we will use the following more general and generic notation. We will denote the available information from source S_i, e.g., a measurement provided by sensor i, as *evidence* e_i. We will denote the desired unknown quantity, i.e., the state of the nature or the parameter, as *hypothesis* h. We will denote the *degree* of *belief* provided for hypothesis h_j given evidence e_i, by $B(h_j/e_i)$. The degree of belief defined here should not be confused with 'the belief function of D-S theory. $B(h_j/e_i)$could be either the posterior probability, if we work in the Bayesian framework, or the belief function of D-S theory, etc. To make the notation simpler, we will write $B_j(e_i)$ for $B(h_j/e_i)$ and, if from the context it is clear that we are addressing a particular hypothesis, say h, we will drop the index and simply write $B(e_i)$. This should cause no confusion.

Now suppose we are given m pieces of evidence, $e_1,...,e_m$. Let $B(e_1,...,e_m)$ represent the *combined* degree of belief for a hypothesis, say h_j, where again for the sake of simplicity of notation, the index j has been dropped. We can write the B function above as

$$B(e_1,...,e_m) = G_m(B(e_1),...,B(e_m)) = G_m(b_1,...,b_m) \tag{4.2.1}$$

where $b_i = B(e_i)$, is the degree of belief provided by the individual source i. The Function G, must have certain nice properties. For instance, the ordering of its argument should not change its value. Furthermore, if we can find another function g such that

$$G_m(b_1,...,b_m) = g(G_{m-1}(b_1,...,b_{m-1}),b_m) \tag{4.2.2}$$

Following cheng et. al. (1988), we will call G_m *binary decomposable* and we will call the function g its *binary operator.* What this basically says is that, we can obtain the combined degree of belief by taking any two pair of evidences, get their joint degree of belief and combine that number with the third piece of evidence, etc. It should be obvious that computing the overall degree of belief in this fashion inherently assumes some type of "independence" among various pieces of evidence. This will become clearer shortly. Now, we will list a set of properties that we would expect any reasonable combination rule to have.

Property p l) *Commutativity*

$$g(a,b)=g(b,a) \quad \text{for all } a, b. \tag{4.2.3}$$

Property p2) Associativity

$$g(g(a,b),c)=g(a,g(b,c)) \quad \text{for all } a, b. \tag{4.2.4}$$

These two properties imply that pieces of evidence are exchangeable and the order of combination is irrelevant.

Property p3) *Monotonicity*

$$a \le b \text{ implies } g(a,c) \le g(b,c) \quad \text{for all } c. \tag{4.2.5}$$

This property implies that if a piece of evidence is replaced by a stronger one, the combined belief should also be stronger.

Property p4) *Continuity*

For any $a,b,c,$ if $g(a,c) \le u \le g(b,c),$ then there exists d such that $a \le d \le b$ and $g(d,c)=u.$

This property conforms with our human intuition that our combined degree of belief should not change abruptly with a slight change in strength of any pieces of evidence.

An element ∎ that has the property $g(a,I)=a$ for all a, is called the ***identity*** for the binary operator g. An element z that has 'the property $g(a,z)=z$ for all a, is called the annihilator for the binary operator g. Since, we assume commutativity and associativity, the identity and the annihilator are unique, if they exist. The intuitive interpretation for an identity ∎ is that the corresponding source (or piece of evidence) is non-informative and the combined information is solely due to the other source (or piece of evidence). Similarly, an annihilator z, represents piece of evidence so strong that overcomes the information provided by the other source. Typical values for ∎ and z, when they exist and when the belief interval is [0 1] is either the endpoints (0 or 1) or the mid-point (0.5).

It is interesting to note that Abel (1926) and Aczel (1949) were able to show that the solution to the functional equation given by the associative property p2) above that has also commutativity, monotinicity and continuity properties is given b)

$$g(a,b)=h^{-1}(h(a)+h(b)) \tag{4.2.6}$$

where his a continuos and strictly monotone function. As an example, we can consider the following family of operators (called Hamacher's family) indexed by $\gamma,\gamma>0$

$$h(a)=log\left(\frac{\gamma}{1-a}+1-\gamma\right) \tag{4.2.7}$$

with the corresponding binary operator

$$g(a,b)=\frac{a+b+(\gamma-2)ab}{1+(\gamma-1)ab} \tag{4.2.8}$$

Remark:

1) When the range of values for the degree of belief is an interval on the real line, (e.g., which typically is the range [0,1], as opposed to the case where the belief is described in terms of linguistic quantifiers such as {unlikely, likely, very likely, most likely}, etc.), then any binary operator g that satisfies properties p l) - p4) is called a *thread* (Clifford, 1958 and Cheng et. al. 1988). Threads have been studied extensively in the areas of functional equations, measurement theory, etc. A thread that has its endpoints (e.g., 0 and 1 if the range of belief is [0 1]), as its identities is called a *Faucett's thread.* For a comprehensive treatment of threads, see Cheng and Kashyap (1988 and 1989), Aczel (1966), Hajek (1985).

2) A binary operator T, $T : [0,1] \times [0,1] \rightarrow [0,1]$, which has properties p1) - p3), i.e., commutativity, associativity and monotinicity property and has 1 as its identity is also called a *T-norm* and has been studied in statistical metrics context by Menger (1942), and Schweizer and Sklar (1983). Note that general T-norms are not required to have the continuity property. A T-norm that is also continuos and has the additional property that $T(x,x) < x$ for all $x \in (0,1)$ is called an Archimedean T-norm. T-norms have also been investigated in the fuzzy set theory context; see Alsina et. al (1983), and Weber (1983).

Now , we are equipped with the required machinery to examine Bayes' rule and D-S theory for combination of evidence.

4.3 Bayesian Combination Rules

The Bayesian approach to combination of evidence is simple. Given evidence e_1 from source S_1, evidence e_2 form source S_2, etc. regarding hypothesis h_i, where e_1 could be for instance measurement X made with an MSS sensor, e_2 could be measurement Y made with a Radar, etc. , the combined information is given by

$$pr(h_i / e_1 \& e_2 \& ... \& e_m) = \frac{pr(e_1 \& e_2 \& ... \& e_m / h_i) pr(h_i)}{pr(e_1 \& e_2 \& ... \& e_m)} \quad (4.3.1)$$

where above, knowledge of joint behavior of sources under hypothesis h_i is required. This information is usually rarely available. So some sort of (statistical) independence assumptions are needed to be able to proceed any further. Statistical independence has the clear meaning that probability of *conjunction* of events can be written as the *product* of probabilities of the individual events. Common types of statistical independence are:

1) The conditional independence of evidence on atomic hypottieses assumption (CI) :

$$pr(e_1 \& ... \& e_m / h_i) = \prod_{j=1}^{m} pr(e_j / h_i) \quad \text{for } i = 1.2,...,n. \quad (4.3.2)$$

2) The global independence assumption (GI):

$$pr(e_1 \& ... \& e_m) = \prod_{j=1}^{m} pr(e_j) \quad (4.3.3)$$

3) The Conditional independence on the negation of hypotheses assumption (CIN):

$$pr(e_1 \& ... \& e_m / h_i^C) = \prod_{j=1}^{m} pr(e_j / h_i^C) \quad \text{for } i = 1.2,...,n. \quad (4.3.4)$$

where h_i^C is the set - theoretic complement of h_i.

Of course one can make a combination of CI, CIN, and GI assumptions. Note that for $n = 2$ CI and CIN are identical, but for $n > 2$ they are quite different.

Let us now see how the Bayesian approach handles combination of information. The available information here are the sampling distributions $pr(e_j/h_i)$, and the priors $pr(h_i)$, from which we can compute the posterior probabilities of individual sources, $pr(h_i/e_j)$. The combination rule depends on the independence assumptions made. Assuming CI independence, Bayes' rule given in (4.4.1) becomes

$$pr(h_i/e_1\&\ldots,\&e_m) = \frac{(pr(h_i))^{1-m}\prod_{j=1}^{m} pr(h_i/e_j)}{\sum_{k=1}^{n}\left\{(pr(h_k))^{1-m}\prod_{j=1}^{m} pr(h_k/e_j)\right\}} \tag{4.3.5}$$

Under simultaneous CI and GI independence assumptions, Bayes' rule of (4.4.1) becomes

$$pr(h_i/e_1\&\ldots\&e_m) = pr(h_i)\prod_{j=1}^{m}\frac{pr(e_j/h_i)}{pr(e_j)} \tag{4.3.6}$$

$$= (pr(h_i))^{1-m}\prod_{j=1}^{m} pr(h_i/e_j) \tag{4.3.7}$$

This is the rule recommended by Swain et. al, (1985) and is also used in the expert system MYCIN, a medical diagnosis system. clinical consultation program.

Applying both CI and CIN, Bayes' rule of (4.4.1) becomes

$$pr(h_i/e_1\&\ldots,\&e_m)=\frac{(pr(h_i))^{1-m}\prod_{j=1}^{m}pr(h_i/e_j)}{(pr(h_i))^{1-m}\prod_{j=1}^{m}pr(h_i/e_j)+(1-pr(h_i))^{1-m}\prod_{j=1}^{m}(1-pr(h_i/e_j))} \tag{4.3.8}$$

This is the rule used in PROSPECTOR, an expert system for mineral exploration and interpretation of geological data. See Goicoechea (1988), Frybach (1978), and Buxton (1989).

It is important to realize that all of the above variants of Bayes' rule are decomposable. The binary operator for each rule can be easily obtained by setting the number of evidence m=2. For instance, the rule (4.4.4), has binary operator

$$g(pr(h_i/e_1),pr(h_i/e_2))\overset{\Delta}{=}g(p_1,p_2)=pr(h_i/e_1\&e_2)=\frac{p_1.p_2}{p_1.p_2+(1-p_1).(1-p_2)} \tag{4.3.9}$$

Furthermore, the binary operator has 0.5 as the identity, since $g(p_1,0.5)=p_1$, or $g(0.5,p_2)=p_2$. And 0 (and 1) are the annihilators of the binary operator. That is, $g(p_1,0)=0$ for all p_1 except $p_1=1$, or $g(0,p_2)=0$ for all p_2 except $p_2=1$; similarly $g(p_1,1)=1$ for all p_1 except $p_1=0$, and $g(1,p_2)=1$ for all p_2 except $p_2=0$. The interpretation here is that if one piece of evidence rejects (or confirms) a hypothesis with certainty, then as long as the other source does not confirm (or reject) with certainty the same hypothesis, its information is irrelevant. The case where one piece of evidence confirms a given hypothesis with certainty and the other piece rejects the same hypothesis with certainty, i.e. complete contradiction, would lead to an undefined value for the combined belief $g(0,1)$.

We also like to mention that, one can easily verify that the binary operators for each of the above rules have all the desired commutativity, associativity, monotinicity and continuity properties (i.e., properties p1) - p4)).

An important question remaining here is which rule should be used; i.e., what independence assumption(s) must be made? The answer is simple: Ideally, *none!* That are no independence assumptions that must be made, and Bayes' rule of (4.4.1) must be used. This means that if it is possible to obtain the joint distributions without any independence assumptions, one should do so. But in real applications the joint information may not be available. Then, we claim that only conditional independence (CI) alone should be made. One should definitely avoid the combination of (CI), (Gi), or (CIN) independence assumptions. The reason for this discrepancy becomes clear after the following definitions due to Cheng et. al. (1986).

DEFINITION 4.1: Evidence e_j is said to be *irrelevant to* the hypothesis h_i if

$$pr(h_i/e_j) = pr(h_i) \tag{4.3.10}$$

Otherwise, it is said to be relevant to h_i.

DEFINITION **4.2:** Evidence e_j is said to be *completely irrelevant* if it is irrelevant to every hypothesis:

$$pr(h_i/e_j) = pr(h_i) \text{ for all } i = 1,...,n. \tag{4.3.11}$$

The following results due to Glymore (1985), Johnson (1986), Cheng and Kashyap (1986), and Pednault et. al. (1981) show that combination of any two or more of CI, GI and CIN could lead to undesirable consequences. That is

THEOREM 4.1: Under simultaneous CI and CIN assumptions, for each hypothesis h_i there can be *at most one* relevant evidence. Furthermore, at least $max\{0,(m-\lfloor n/2 \rfloor)\}$ pieces of evidence will be *completely* irrelevant.

Similar results can be stated for combination of CI and GI, CIN and GI, etc. The main conclusion here is that CI alone is usually sufficient and no other independence assumptions should be made.

In closing this section, we also like to making the following remarks in support of Bayesian updating rule.

Remarks:

1) Cox's (1946) postulated seven desirable properties, among which were commutativity, associativity, monotonicity and continuity, for any belief updating rule and proceeded to prove that the resulting belief function is a probability. See also Schocken and Kleindorfer (1989).

2) As the amount of information (data or evidence) increases, the uncertainty in the combined belief diminishes; put in different words, asymptotically the combined posterior probability approaches a 0-1 distribution, where the true but unknown hypothesis will have posterior probability of one and the rest will have posterior probability of zero.

3) Note that in decision problems, often we do not need to compute the denominators in the Bayesian combination or updating rule(s) above.

Next, we investigate the D-S combination rule and the independence assumptions made in there. However, since D-S is not the main focus of our thesis, we will not give the full details here. Interested readers can consult the original papers of Demspter (1966, 1967, 1968), Shafer (1976, 1982), Klir (1988), Smets (1981, 1988, 1990) and many other interesting papers written since. A comprehensive list of references is provided in the reference section.

4.4 D-S Combination Rule

Recall from Chapter 3 that in the D-S theory sample space is required to be finite and the belief functions, Bel, and their conjugates plausibility functions, Pl are ∞ - monotone and ∞ - alternating capacities, respectively. Also recall, that for every belief function, there is a unique mobius inverse function, m, called the basic probability assignment (bpa) function. The combination rule can be explained more conveniently in terms of the bpa functions.

The D-S combination rule (also known as Dempster's orthogonal sum) states that given two *entirely distinct* bodies of evidence e_1 and e, with their corresponding bpa functions m, and m, the combined bpa. function m_{12} is expressed by

$$m_{12}(A) = m_1(A) \oplus m_2(A) = \frac{\sum_{B \cap C = A} m_1(B).m_2(C)}{1 - \sum_{B \cap C = \phi} m_1(B).m_2(C)} \qquad (4.4.1)$$

The above combination rule has the following desirable properties:

1) The combination rule is obviously decomposable and m_{12} is the binary operator for the combination rule.

2) The binary operator $m,,$ is commutative and associative.

3) 0 and 1 are the annihilators for m_{12}.

The undesirable properties of the above rule, however, are:

1) The meaning of "entirely distinct" bodies of evidence is not clear. Many researchers have tried to find statistical and other interpretations for this requirement with limited success. See Voorbraak (1991).

2) Many researchers , e.g., Zadeh (1984) and (1986), Lammer (1986), Voorbraak (1991), etc. have constructed examples where using Dempster's rule would lead to inconsistencies. The main problem is the denominator in (4.5.1) which serves as *re-normalization* factor.

3) We showed in Chapter 3, that the conditioning rule that follows directly from Dempster's combination rule is inconsistent with the desired rule.

4) D-S has been used for combination of statistical evidence and prior evidence. Shafer (1982), Walley (1987), Kim (1990). But Walley (1987) has elaborately proven that Dempster's rule is not generally suitable for combing evidence from independent observations nor it is suitable to combine prior belief with observation evidence.

5) The number of computations required in Dempster's rule grows exponentially; Orponen (1990), Kennes (1992). This is mainly due to the fact that D-S theory works with the power sets. To be exact, if the sample space Ω has n elements, to compute (4.5.1), we need to perform $(2^{2^n} - 2^n)$ additions and 2^{2^n} multiplications. And to compute the Bel function (or the Pl function) we need to do $(3^n - 2^n)$ extra additions. Ignoring the required addition operations, this implies that to combine two sources in D-S theory, the time complexity is of order $O(2^{2^n})$. And if there are K sources, then the time complexity is $O(2^{K^n})$.

6) Furthermore, in D-s Theory one needs to specify the values of bpa's on the power set, whereas in probability theory one needs to specify probability density (or actually mass) function only on the sample space. that is if the sample space Ω is finite and has n elements, one needs to specify values of 2^n basic probabilities, opposed to specifying n values for the probability mass function. So, if one has difficulty in specifying the probability mass function, it is not clear how specifying bpa function would be any easier! Also, in terms of storage, above implies that D-S requires exponentially more storage space. In the next Chapter, we will examine Bayesian approaches for combination of imprecise information and provide their computational complexities and

highlight the savings they offer in terms of storage and computational complexities.

CHAPTER 5

COMBINATION OF IMPRECISE SOURCES OF INFORMATION

5.1 Introduction

In this Chapter we will introduce three different approaches for combination of imprecise sampling distributions (possibly from multiple sources) and imprecise priors. Throughout the Chapter, we will adhere to Bayes' rule (or its new version for Capacities). In the case of multiple sources, we will only make a conditional independence (CI) assumption to combine information. We will consider the computational burden of each method and compare them with the computational complexity cost in D-S theory.

In the first approach, we utilize the *extreme point* representation suggested by White (1986) and obtain the *posterior* extreme points from the extreme points of the priors and the extreme points of the sampling densities. We look at the computational complexity of this approach and compare it to the computational complexity of D-S Theory. Even though this approach has better computational complexity than D-S approach and does not suffer from some of the major criticisms of D-S Theory, its computational complexity may still be a problem. Thus we propose a second method that uses a linearization technique of Wasserman, Lavin and Wolpert (1993). This approach is iterative and converts a nonlinear optimization problem for finding upper (and lower) posteriors or posterior related quantities into a sequence of simpler linear optimizations. We provide several examples here. As the third and final approach, we look at the product or the joint space of measurements and parameters, $\chi \times \Theta$. We realize that *if* the class of joint distributions (or densities) can be described in terms of

joint 2-Choquet capacities, then we can utilize our Theorem of Chapter 3 and find the *posterior* capacities directly. This approach has the smallest computational complexity. We provide several examples.

5.2 Extreme Point Representation

Let us assume that the measurement space and the parameter space are both finite. Furthermore, let us assume that the set of imprecise models and priors are *Convex sets.* This is a relatively mild requirement arid many useful neighborhoods are convex. For example, when the imprecisiori is described in terms of linear inequalities, the resulting set is convex. A description of imprecision in terms of linear inequalities is often very natural and practical. Below, we provide three cases to motivate the idea. Case 1 corresponds to the situation where the available information translates into a general convex set. The idea is explained by a typical example from medicine. Case 2 corresponds to imprecision specified by general lower and upper bounds. Case 3 corresponds to an important special case of Case 2 which has a very natural interpretation and the lowest computation cost; i.e., that of point valued or precise probabilities. We begin with Case 1 with an example which is due to White (1986).

CASE 1: Here we will assume that both the parameter space, Θ, and the measurement space, X, are finite.

Consider a patient with joint pain who is assumed to be in one of the four following mutually exclusive states of health:

1) fibrositis $\stackrel{A}{=} \theta_1$
2) cervical nerve compression $\stackrel{A}{=} \theta_2$
3) polymyalgia rheumatica $\stackrel{A}{=} \theta_3$
4) nonspecific joint pain $\stackrel{A}{=} \theta_4$

Assume that a physician makes the following statements, based on the patient's history

1) The likelihood that the patient has nonspecific joint pain is between 2.0 to 2.5 times that the patient has fibrositis.
2) The likelihood that the patient has cervical nerve compression is nine to ten times the likelihood that the patient has polymyalgia rheumatica.
3) The likelihood that the patient has cervical nerve connpression is five times as great as the likelihood that the patient has nonspecific joint pain.

That is:

$$2.0\pi(\theta_1) \le \pi(\theta_4) \le 2.5\pi(\theta_1)$$
$$9\pi(\theta_3) \le \pi(\theta_2) \le 10\pi(\theta_3)$$
$$\pi(\theta_2) = 5\pi(\theta_4)$$

Note that the above information corresponds to the following set of priors:

$$\Gamma = \left\{ \begin{array}{l} \pi : 2.0\pi(\theta_1) \le \pi(\theta_4) \le 2.5\pi(\theta_1);\ 9\pi(\theta_3) \le \pi(\theta_2) \le 10\pi(\theta_3); \\ \qquad \pi(\theta_2) = 5\pi(\theta_4);\ \sum_{i=1}^{4} \pi(\theta_i) = 1 \end{array} \right\} \quad (5.2.1)$$

Suppose that the models $f(x/\theta_i)$, $i = 1,...,4$ for the above disorders are also known partially. More specifically, assume that the physician determines that there are trigger points (with or without modules) in the soft tissue surrounding the affected area and can only state the following

1) The likelihood that trigger points will be found is the same as if the patient has cervical nerve compression or if the patient has nonspecific joint pain.
2) The likelihood that trigger points will be found if the patient has cervical nerve compression is between one and two times the

likelihood that trigger points will be found if the patient has polymyalgia rheumatica.

3) The likelihood that trigger point will be found if the patient has fibrositis is 7 to 8 times the likelihood that trigger points will be found if the patient has cervical nerve compression.
4) The probability that trigger points will be found in a patient with polymyalgia rheumatica is at least 0.01.
5) The probability that trigger points will be found in a patient with fibrositis is between 0.90 and 0.95.

Let x represents the result of the physicians measurement (or examination), where x can have only two possible values of trigger points being present or absent. Then the above information can be summarized as:

$$f(x = \textbf{trig. pts. } found \mid \theta_2) = f(x = \textbf{trig. pts. } found \mid \theta_4)$$
$$f(x = \textbf{trig. pts. } found \mid 8,) \le f(x = \textbf{trig. pts. } found \mid \theta_2) \le 2\, f(x = \textbf{trig. pts. } found \mid 8,)$$
$$7\, f(x = \textbf{trig. pts. } found \mid \theta_2) \le f(x = \textbf{trig. pts. } found \mid 8,) \le 8 f(x = trig.\ pts.\ found \mid \theta_2)$$
$$f(x = \textbf{trig. pts. } found \mid \theta_3) \ge 0.01$$
$$0.90 \le f(x = \textbf{trig. pts. } found \mid 8,) \le 0.95$$

(5.2.2)

Now, suppose the physician examines a patient and detects the presence (or absence) of the trigger points. Given the above measurement and the imprecise information (5.2.1) and (5.2.2) regarding the priors and the likelihoods or the models, we like to determine the set of posterior probabilities that the patient has any of the given disorders.

Note that the set of priors specified in (5.2.1) is a convex set with finite number of extreme points. Let ε_{priors} denote the set of prior extreme points and $\underline{\pi}^{(m)}$ denote its elements.

Note also that any prior in the set Γ can be expressed as a linear convex combination of the above extreme points.

Similarly, given the observation that a trigger point was found, let $\Gamma_{likelihoods}$ represent the set of possible likelihoods given by (5.2.2). Let $\varepsilon_{likelihoods}$ be the set of extreme points of $\Gamma_{likelihoods}$ and $\underline{f}^{(n)}$ be its elements.

Then the set of posterior extreme points $\varepsilon_{posteriors}$, with elements; $\underline{\pi}^{(k)}(.|.)$ can be computed by using the classical Bayes' rule which in the vector form can be written as

$$\underline{\pi}^{(k)}(./x = trig.\ pts.\text{found}) = \begin{bmatrix} \pi^{(k)}(\theta_1/x = trig.\ pts.\text{found}) \\ . \\ . \\ . \\ \pi^{(k)}(\theta_4/x = trig.\ pts.\text{found}) \end{bmatrix} \tag{5.2.3}$$

$$= \frac{\underline{\pi}^{(m)} \otimes \underline{f}^{(n)}}{\sum\left(\underline{\pi}^{(m)} \otimes \underline{f}^{(n)}\right)} \tag{5.2.4}$$

where $\underline{\pi}^{(m)} \otimes \underline{f}^{(n)}$ is the vector whose *jth* element is the product of *jth* element of the vector $\underline{\pi}^{(m)}$ and the *jth* element of the vector $\underline{f}^{(n)}$, and $\sum\left(\underline{\pi}^{(m)} \otimes \underline{f}^{(n)}\right)$ represent the sum of the elements of the vector $\underline{\pi}^{(m)} \otimes \underline{f}^{(n)}$.

If we let Q represent the set of posteriors obtained by applying the Bayes' rule to the *entire* set of priors and the *entire* set of likelihoods (and not just the extreme points), and let $CH(A)$ denote the convex hull of set A, we have the following useful results proofs of which can be found in White (1986).

Result 1: $\varepsilon_{posteriors} \subseteq Q \subseteq CH(\varepsilon_{posteriors})$

Result 2: Let *ext* represent either minimum or maximum, and $C \in R^n$ where n is the dimension of the parameter space Θ. Then

$$ext\left\{yc:\ y\in\varepsilon_{posterior}\right\}=ext\left\{yc:\ y\in Q\right\}=ext\left\{yc:\ y\in CH\left(\varepsilon_{posterior}\right)\right\}. \qquad (5.2.5)$$

That is, for instance, to find the minimum (or the maximum) posterior probability of θ_i, i.e., $ext\{y_i:\ y\in Q\}$, one only needs to search through the posterior extreme points for the minimum (or the maximum) value.

For the above example, the upper and lower posterior probabilities using eq. (5.2.4) and result 2 (eq. (5.2.5)) can easily be computed as

$$\begin{aligned}
&0.253\le\pi(\theta_1\mid \textit{trig. pts. found})\le 0.443\\
&0.418\le\pi(\theta_2\mid \textit{trig. pts. found})\le 0.607\\
&0.018\ I\ \pi(\theta_3\mid \textit{trig. pts. found})\ 1\ 0.077\\
&0.083\le\pi(\theta_4\mid \textit{trig. pts. found})\le 0.125
\end{aligned} \qquad (5.2.6)$$

If the set of priors and the set of likelihoods have N_{priors} and $N_{likelihoods}$ extreme points, respectively, then to compute the set of all posteriors extreme points, we need to perform $3(n+1)*\binom{N_{total}}{2}$ multiplications and $n*\binom{N_{total}}{2}$ additions, where $N_{total}=N_{prior}+N_{likelihoods}$, and n is the dimension of the parameter space,

Of course the above approach can be easily extended to several sources using the! Bayes' rule and any of the independence assumptions, in particular, the conditional independence (CI) assumption. Given there are S sources and if we assume the set of likelihoods corresponding to source k has $N^k_{likelihood}$ extreme points, then to compute the posterior extreme points we need to perform $(n+1)*S*\binom{N_{total}}{S}$ multiplications, and $n*\binom{N_{total}}{S}$ additions, where

$N_{total} = N_{prior} + \sum_{i=1}^{S-1} N^{i}_{likelihood}$. Since $\binom{n}{m} = O\left(\frac{n^m}{m!}\right)$, the computation cost of above procedure is of order $O\left((n+1)\frac{(N_{total})^{S+1}}{(S-1)!}\right)$.

Of course, the above computational costs do not include the cost of determining the extreme points for the priors and the likelihoods, which in general is a nontrivial task.

Even though, there is no Dempster-Shafer interpretation for the scenario presented in the Case 1, and the only statistical interpretation given by Shafer (1982) corresponds to the case of combination of precise likelihoods with imprecise priors, described by belief and plausibility functions, we like to examine the computational complexity of the Dempster-Shafer rule to obtain a feeling for the number of computations involved in the different approaches.

Recall the worst case computational complexity of Demspter-Shafer combination rule [Kennes (1992), Henkind and Harrison (1988), Orponen (1990)] for combining two sources, where the parameter space has n elements, involves 2^{2n} multiplications and $2^n(2^n-1)$ additions; i.e., is of order $O(2^{2n})$ computation. By induction the worst case computational complexity of Dernpster-Shafer rule to combine S sources is $O(2^{Sn})$.

Direct comparison of computational complexities of the Dempster-Shafer rule and the extreme point approach is not possible since the later depends on the total number of extreme points, N_{total}, and there is no closed form expression for this quantity. However, unless N_{total} is $O(2^n)$ or larger, the extreme points approach would be more efficient in terms of computational complexity.

CASE 2: Here we will assume that the parameter space Θ is finite, but the measurement space X could be either finite or continuos.

In many practical situations, the available imprecise information can be expressed in terms of upper and lower bounds for the unknown quantity. A typical example is using confidence interval estimates.

More precisely, let us consider the imprecise priors and represent the available information as

$$\begin{aligned} &\ell(\theta_1) \le \pi(\theta_1) \le u(\theta_1) \\ &\ell(\theta_2) \le \pi(\theta_2) \le u(\theta_2) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.2.7) \\ &\ell(\theta_n) \le \pi(\theta_n) \le u(\theta_n) \\ &\sum_{i=1}^{n} \pi(\theta_i) = 1 \end{aligned}$$

Recall from Chapter 2 that, for the set of imprecise priors Γ defined by (5.2.7) to be non empty, we need the following simple requirement

$$R1) \quad \sum_{i=1}^{n} \ell(\theta_i) \le 1 \text{ and } \sum_{i=1}^{n} u(\theta_i) \ge 1. \qquad (5.2.8)$$

Furthermore, for this set not to be "unnecessarily too large", we require

$$R2) \quad \begin{cases} \ell(\theta_i) \ge 1 - \sum_{\substack{j=1 \\ j \neq i}}^{n} u(\theta_j) \\ u(\theta_i) \le 1 - \sum_{\substack{j=1 \\ j \neq i}}^{n} \ell(\theta_j) \end{cases} \qquad (5.2.9)$$

As we mentioned earlier, there are many ways that one could come up with the lower and upper bounds above. They represent our minimum and maximum prior beliefs in occurrence of various outcomes. Lower and upper bounds for the likelihoods (or the sampling densities) can come about, for example, when

they are estimated from small size training data and are expressed as lying within pairs of confidence limits.

Given a set of linear inequalities such as (5.2.7), if the set is non empty but too large, we can refine the bounds to get a set that satisfies requirement R2 as

$$\ell'(\theta_j) = max\{\ \ell(\theta_j)\ ,\ 1 - \sum_{\substack{i=1 \\ i \neq j}} u(\theta_i)\ \} \tag{5.2.10}$$

$$u'(\theta_j) = min\{\ u(\theta_j)\ ,\ 1 - \sum_{\substack{i=1 \\ i \neq j}}^{n} \ell(\theta_i)\ \} \tag{5.2.11}$$

Here, the resulting set of priors is not only convex, but also a polytope which again is completely determined in terms of its finite number of extreme points. Extreme points of a convex polytope can be found using different methods such as linear programming, etc. See Balinski (1961), Matheiss ancl Rubin (1980), Karmarker (1984), Ho and Kashyap (1965). In general, for an arbitrary convex polytope the task of finding all its extreme points is usually nontrivial. But due to the simple structure present in our representation, it is easy to see that in an n-dimensional space, the corresponding convex polytope could have minimum of n and maximum of $n(n-1)$ vertices, and those can be computed relatively easily.

Given an observation $x = x_o$, to combine the information provided by the lower ancl upper bounds for the priors and the likelihoods using the Bayes' rule, in the worst case we need to perform $O\left(2^{2n-1}\right)$ multiplications. See Figure 5.1 below. This is because in the worst case there would be $n(n-1)$ extreme points in the prior convex polytope, and 2^n possible combinations for the extreme likelihood values, thus the overall

$$Worst\ Case\ Cost = \binom{n(n-1)+2^n}{2} = 2^{2n-1} + (n^2 - n)2^n - 2^{n-1} + \frac{n^4}{2} + \frac{n}{2} = O\left(2^{2n-1}\right)$$

(5.2.12)

which is comparable to the combination cost of Dempster-Shafer rule and can be easily generalized for S sources.

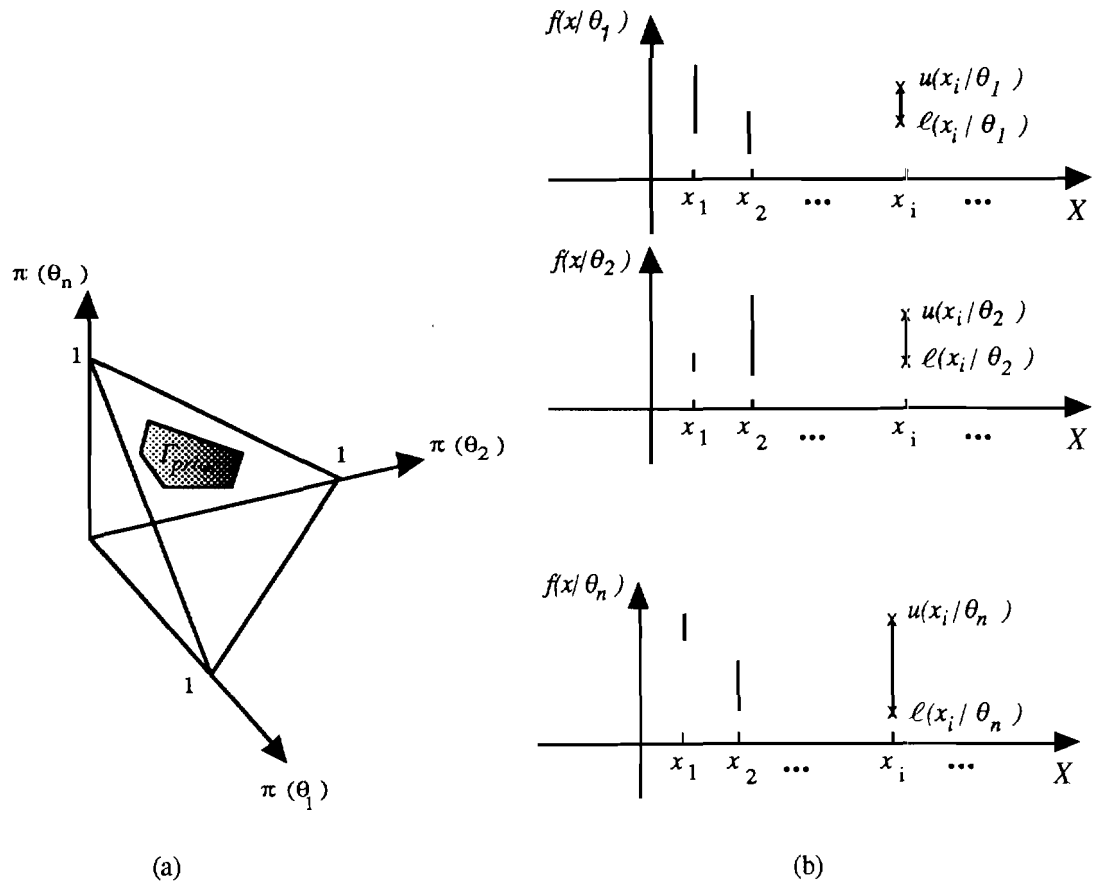

Figure 5.1 - (a) The Set of Imprecise Priors; (b) Upper and Lower Bounds Specification of the Imprecise Sampling Probability Mass Functions.

CASE 3: Here we will assume that both the parameter space Θ and the measurement space X are finite. This is an important special case of Case 2 where we can specify only the lower bounds; i.e., the minimum degrees of beliiefs, and we let the upper bounds to be the *largest* values allowed by the requirement R2. That is, considering imprecise priors, we have

$$\Gamma_{prior} = \left\{ \begin{array}{c} \ell(\theta_i) \le \pi(\theta_i) \le u(\theta_i)\ ;\ i = 1, \ldots, n \\ \ell(\theta_i) \ge 0\ ,\ u(\theta_i) = 1 - \sum_{\substack{j=1 \\ j \ne i}}^{n} \ell(\theta_j)\ ;\ i = 1, \ldots, n \end{array} \right\} \qquad (5.2.13)$$

Note that

$$u(\theta_i) - \ell(\theta_i) = 1 - \sum_{j=1}^{n} \ell(\theta_j) \qquad (5.2.14)$$

which is independent of the index i; i.e., the *range* of uncertainty specified by the width of the interval is the same for all $\theta_i's$. That is the upper bounds are, in a manner of speaking, non-informative.

Similarly for the imprecise likelihoods, we have

$$\Gamma_{likelihood}^{\theta_i} = \left\{ \begin{array}{l} \ell(x_j/\theta_i) \le f(x_j/\theta_i) \le u(x_j/\theta_i)\ ;\ j = 1, 2, \ldots, M \\ \ell(x_j/\theta_i) \ge 0 \ \textbf{and}\ u(x_j/\theta_i) = 1 - \sum_{\substack{k=1 \\ k \ne j}}^{M} \ell(x_k/\theta_i) \end{array} \right\} \qquad (5.2.15)$$

Since in this section, discussion regarding the priors and the likelihoods are almost identical, we will use the generic notation

$$\Gamma_{\ell,u} = \left\{ \begin{array}{c} \ell(z_i) \le p(z_i) \le u(z_i) \\ \ell(z_i) \ge 0 \ \textbf{and}\ u(z_i) = 1 - \sum_{k \ne i} \ell(z_k)\ ;\ z_i \in Z \end{array} \right\} \qquad (5.2.16)$$

to refer to the set of priors or the set of likelihoods. Above, subscript ℓ,u is used to emphasize the role of both the upper and the 'lower bounds.

Let T denote the set of *all* possible probability distributions. Since we are considering finite spaces, T is simply the (appropriate) probability simplex. Let $\mathcal{D} \subseteq \mathcal{P}$ be an arbitrary set of probability distributions that satisfies the requirements *R1* and *R2.* Let

$$\ell(z_i) \stackrel{\Delta}{=} \inf_{p \in D} p(z_i) \; ; \; z_i \in Z \tag{5.2.17}$$

be the lower bound of $\mathcal{P}$, at point z_i, and Let

$$\Gamma_\ell \stackrel{\Delta}{=} \{ \, p \in \mathcal{P} : \ell(z_i) \le p(z_i) \; ; \; z_i \in Z \, \} \tag{5.2.18}$$

be the set of all probability distributions that are larger than the lower bound at every point, $z_i \in Z$.

Note that, in general, $\Gamma_\ell \neq \mathcal{P}$. When $\Gamma_\ell = \mathcal{P}$, we say Γ_ℓ and ℓ are *representable* (pointwise) and write them as (Γ_ℓ, ℓ) to contrast them with the more general notion of representability defined in Chapter 2, Section 3. Figure 5.2 shows an example of the above idea in 3-dimensions. Note that, in 3-D the set Γ_ℓ is an equilateral triangle.

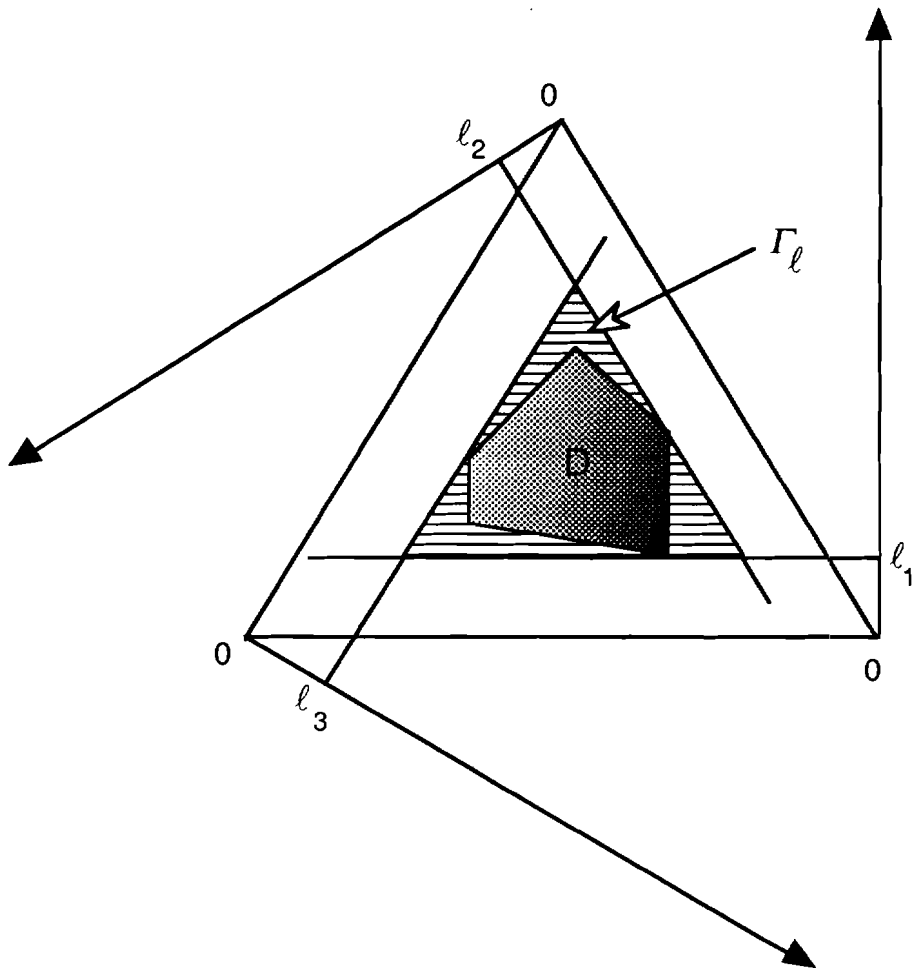

Figure 5.2 - An Arbitrary Set of Imprecise Probabilities, D, and Its Corresponding Set Γ_ℓ.

Similarly, let $u(z_i) = 1 - \sum_{k \neq i} \ell(z_k)$ and define the set of probability distributions specified *only* by the upper bound as

$$\Gamma_u \stackrel{\Delta}{=} \{\, p \in \mathcal{P} :\ p(z_i) \leq u(z_i)\ ;\ z_i \in Z \,\} \tag{5.2.19}$$

Note that, in general, $\Gamma_\ell = \Gamma_{\ell,u} \subseteq \Gamma_u$. Proof is simple and is omitted. This implies that specification by the upper bounds alone is not enough, and we need to consider $\Gamma_{\ell,u}$, or Γ_ℓ. Figure 5.3 indicates this idea in the case of 3-dimensions.

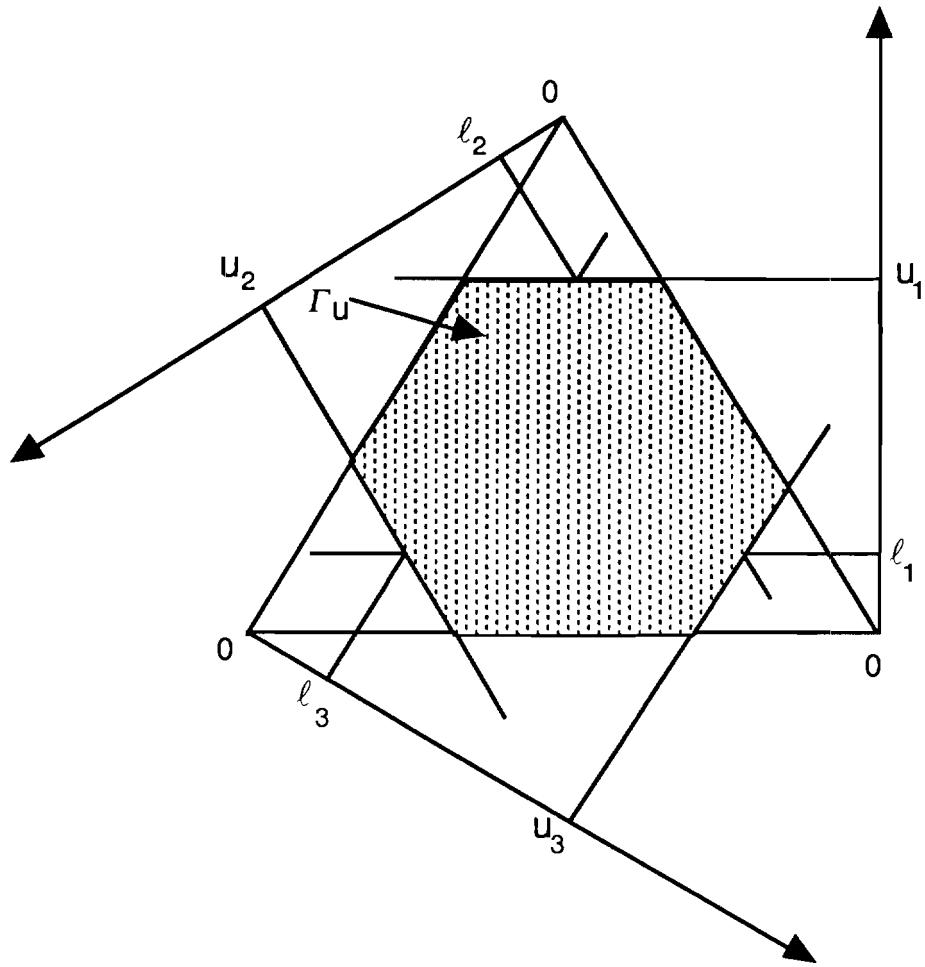

Figure 5.3 - An Arbitrary Set of Imprecise Probabilities', D, and Its Corresponding Set Γ_u.

Now, let $\Gamma^{\ell,u}_{priors}$ and $\left\{\Gamma^{\ell,u}_{\theta_i}\ ;\ i=1,\ldots,n\right\}$ represent the set of imprecise priors and the imprecise sampling distributions, respectively. That is

$$\Gamma^{\ell,u}_{priors}=\left\{\begin{matrix}\pi:\ \ell(\theta_i)\le\pi(\theta_i)\le u(\theta_i)\ ;\ i=1,\ldots,n\\ u(\theta_i)=1-\sum_{k\,t\,i}\ell(\theta_k)\end{matrix}\right\} \qquad (5.2.20)$$

and

$$\Gamma_{\theta_i}^{\ell,u} = \left\{ \begin{array}{c} f(x_j/\theta_i) : \ell(x_j/\theta_i) \le f(x_j/\theta_i) \le u(x_j/\theta_i) \; ; \; j = 1, \ldots, M \\ u(x_j/\theta_i) = 1 - \sum_{k \neq j} \ell(x_k/\theta_i) \end{array} \right\}, \quad i = 1, \ldots, n. \tag{5.2.21}$$

The following theorem shows how we can obtain the lower and upper bounds for the resulting set of imprecise posteriors.

THEOREM: *Given an observation $x = x_j$, and the above representation for the imprecise priors and the imprecise sampling distributions, then*

$$\ell(\theta_i/x_j) \overset{\Delta}{=} \inf_{\pi \in \Gamma_{priors}^{\ell,u}} \inf_{\substack{f(./\theta_i) \in \Gamma_{\theta_i}^{\ell,u} \\ i=1,\ldots,n}} \frac{f(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i)} \tag{5.2.22}$$

$$= \frac{\ell(x_j/\theta_i)\,\ell(\theta_i)}{\ell(x_j/\theta_i)\,\ell(\theta_i) + u(x_j/\theta^*)\,u(\theta^*) + \sum_{\substack{k=1 \\ k \neq i,* }}^{n} u(x_j/\theta_k)\,\ell(\theta_k)} \tag{5.2.23}$$

where $\theta^* = arg \left\{ \max_{\substack{k \\ k \neq i}} u(x_j/\theta_k) \right\}$, *provided in (5.2.23) we are not dividing by zero.*

The upper posterior probabilities can be computed as

$$u(\theta_i/x_j) = \frac{u(x_j/\theta_i)\,u(\theta_i)}{u(x_j/\theta_i)\,u(\theta_i) + \sum_{\substack{k=1 \\ k \neq i}}^{n} \ell(x_j/\theta_k)\,\ell(\theta_k)} \tag{5.2.24}$$

Proof: See appendix A.2.

Let us consider the following simple example to illustrate the method. Suppose the parameter space $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and the measurement space $\mathcal{X} = \{x_1, x_2, x_3\}$.

Furthermore, the available information is expressed as lower bounds for the priors and the conditionals as

$$0.1 \le \pi(\theta_1); \quad 0.4 \le \pi(\theta_2); \quad 0.2 \le \pi(\theta_3); \tag{5.2.25}$$

$$\begin{cases} 0.3 \le f(x_1 / \theta_1); \\ 0.2 \le f(x_2 / \theta_1); \\ 0.0 \le f(x_3 / \theta_1); \end{cases} \quad \begin{cases} 0.2 \le f(x_1 / \theta_2); \\ 0.4 \le f(x_2 / \theta_2); \\ 0.2 \le f(x_3 / \theta_2); \end{cases} \quad \begin{cases} 0.5 \le f(x_1 / \theta_3); \\ 0.1 \le f(x_2 / \theta_3); \\ 0.1 \le f(x_3 / \theta_3); \end{cases} \tag{5.2.26}$$

Given the observation $x = x_2$, posterior lower and upper probabilities can be found using (5.2.23) and (5.2.24) as

$$\begin{cases} 0.039 \; I \; \pi(\theta_1 / x_2) \le 0.609; \\ 0.308 \le \pi(\theta_2 / x_2) \, 10.913; \\ 0.037 \; I \; \pi(\theta_3 / x_2) \, 50.526. \end{cases} \tag{5.2.27}$$

Above method can easily be extended for combination of information from multiple sources. For sake of simplicity of notation and without loss of generality we will consider combination of only two sources. Let us denote the observations from source 1 and source 2, with discrete random variables X and Y, respectively. Furthermore, let us assume that available information regarding each source can be expressed as lower bounds on the conditionals; i.e.,

Source 1: $$\left\{ \ell_{X/\theta}(./\theta_i) \le f_{X/\theta}(./\theta_i) \; ; i = 1, \ldots, n \right\} \tag{5.2.28}$$

Source 2: $$\left\{ \ell_{Y/\theta}(./\theta_i) \le f_{Y/\theta}(./\theta_i) \; ; i = 1, \ldots, n \right\} \tag{5.2.29}$$

and the available information regarding the priors is expressed as

Priors: $$\left\{ \ell(\theta_i) \le \pi(\theta_i) \; ; i = 1, \ldots, n \right\} \tag{5.2.30}$$

Then, under the assumption of Conditional Independence (CI) of the sources, the combined lower posterior probabilities can be computed as

$$\ell(\theta_i / x_j ; y_k) = \frac{N}{D_1 + D_2 + D_3} \tag{5.2.31}$$

where

$$N = \ell_{X/\theta}(x_j/\theta_i)\, \ell_{Y/\theta}(y_k/\theta_i)\, \ell(\theta_i)$$
$$D_1 = N = \ell_{X/\theta}(x_j/\theta_i)\, \ell_{Y/\theta}(y_k/\theta_i)\, \ell(\theta_i)$$
$$D_2 = u_{X/\theta}(x_j/\theta^*)\, u_{Y/\theta}(y_k/\theta^*)\, u(\theta^*)$$
$$D_3 = \sum_{\substack{l=1 \\ l \neq i,^*}}^{n} u_{X/\theta}(x_j/\theta_l)\, u_{Y/\theta}(y_k/\theta_l)\, \ell(\theta_k)$$

and

$$\theta^* = arg \max_{\substack{i \\ i \neq l}} \left\{ u_{X/\theta}(x_j/\theta_i)\, u_{Y/\theta}(y_k/\theta_i) \right\}. \tag{5.2.32}$$

Similar results can be stated foot the combined upper posterior probabilities Proof of (5.2.31) can be found in the appendix A.3.

It is interesting to note that, apart from computing θ^*, the computational complexity of the above method is identical to the computational complexity of the combination of CI sources using the Bayes' rule with the *point-valued* probabilities mentioned earlier. That is, no extra computational cost is involved due to presence of imprecision r uncertainty in the available information.

5.3 Linearization (Iterative) Method

Let Γ^{Π} denote the set of imprecise prior distributions and $\{ f(x/\theta); \theta \in \Theta \}$ be the set of imprecise sampling densities. Let

$$\rho(\phi, \Pi, f) = \frac{\int_{\Theta} \phi(\theta)\, f(x/\theta)\, \Pi(d\theta)}{\int_{\Theta} f(x/\theta)\, \Pi(d\theta)} \tag{5.3.1}$$

represent the posterior quantity of interest. Note that for the following choices of $\phi(\theta)$:

1) $\phi(\theta) = \theta$
2) $\phi(\theta) = I_B(\theta)$
3) $\phi(\theta) = \lambda(\theta, \delta(x))$

we have (1) the posterior mean, (2) the posterior probability of set B, and (3) the posterior expected loss of decision $\delta(x)$. Furthermore, since the priors and the sampling densities are not known precisely, for a given observation or measurement x, we are interested in computing

$$\overline{\rho} \stackrel{\Delta}{=} \sup_{\Pi \in \Gamma^{\Pi}, \{f(x/\theta) \in \Gamma_{\theta}^{f}\}} \rho(\phi, \Pi, f) \tag{5.3.2}$$

$$\underline{\rho} \stackrel{\Delta}{=} \inf_{\Pi \in \Gamma^{\Pi}, \{f(x/\theta) \in \Gamma_{\theta}^{f}\}} \rho(\phi, \Pi, f) \tag{5.3.3}$$

The range of the interval $[\underline{\rho}, \overline{\rho}]$ indicates the degree of robustness of the posterior quantity p to the variations or indeterminacies in the priors and the sarnpling densities. Computation of $\overline{\rho}$ (or $\underline{\rho}$) is complicated by the fact that the above optimization problem is *nonlinear* in Π and $f(x/\theta)$. Fortunately, the following linearization result due to Lavine (1991b), DeRobertis (1978), and Wasserman et. al. (1993) can be used to convert a single nonlinear optimization into a set of simpler linear optimizations.

Theorem: (Linearization) *Let q be any real number and define*

$$C(q) \stackrel{\Delta}{=} \int_{\Theta} (\phi(\theta) - q) f(x/\theta) \Pi(d\theta) \tag{5.3.4}$$

and

$$\overline{C}(q) \stackrel{\Delta}{=} \sup_{\Pi \in \Gamma^{\Pi}, \{f(x/\theta) \in \Gamma_{\theta}^{f}\}} \int_{\Theta} (\phi(\theta) - q) f(x/\theta) \Pi(d\theta) . \tag{5.3.5}$$

Then, $\overline{\rho} > q$ iff $\overline{C}(q) > 0$. A similar result holds for the lower posterior bound $\underline{\rho}$.

Note that $C(q)$ is a linear function of both Π and $f(x/\theta)$. That is, to compute $\bar{\rho}$ we do the following iterative procedure:

1) choose some arbitrary number q.

2) Compute $\bar{C}(q)$.

3) If $\bar{C}(q) > 0$, then $\bar{\rho} > q$. So, we choose another number *larger* than q and go to step 2);
If $\bar{C}(q) < 0$, choose a number *smaller* than q and go to step 2);
if $\bar{C}(q) = 0$, $\bar{\rho} = q$ and stop.

A simple way to implement the above algorithm is to compute $\bar{C}(q)$ over a grid of points $\{q_1, ..., q_k\}$ and then solve $\bar{C}(q) = 0$ numerically.

It is also important to note that usually the set of imprecise priors and imprecise sampling densities, e.g., ε - contamination or band models, are convex sets with easily identifiable extreme points; Berger (1990). Furthermore, as mentioned earlier $C(q)$ is a linear function of Π and $f(x/\theta)$. It is a well known fact that linear functionals over convex constraint sets attain their minimum or maximum at the extreme points of the constraint sets. That is, if we denote the set of extreme points of the imprecise prior set, Γ^{Π}, as $\mathcal{E}^{\Pi}$ and the imprecise sampling densities, $\{\Gamma_\theta^f ; \theta \in \Theta\}$, as $\{\mathcal{E}_\theta^f ; \theta \in \Theta\}$, then

$$\sup_{\Pi \in \Gamma^{\Pi}} \sup_{\{f(x/\theta) \in \Gamma_\theta^f\}} C(q) = \sup_{\Pi \in \mathcal{E}^{\Pi}} \sup_{\{f(x/\theta) \in \mathcal{E}_\theta^f\}} C(q) \qquad (5.3.6)$$

As an application of the above result, let us consider the following example.

Example 5.3.1: Let the imprecision regarding the priors and the models both be described by band models as

$$\Gamma^{\Pi} = \{ \pi : \ell(\theta) \le \pi(\theta) \le u(\theta) \}$$

and

$$\Gamma_\theta^f = \{ f(./\theta):\ \ell(x/\theta) \le f(x/\theta) \le u(x/\theta) \}$$

Band models are useful because they do not require knowledge of the shape of the distribution or nominal model information and allow a wide range of distributions.

In the above example, $\overline{C}(q)$ can be computed easily as:

1) Choose a real number q and fix a $\Pi \in \Gamma^{\Pi}$ (or actually $\Pi \in E^{\bullet\bullet}$).

2) Compute the maximum over the sampling densities: clearly the maximum occurs at

$$f^*(x/\theta) = \begin{cases} u(x/\theta) & \theta \in \{ \theta: (\phi(\theta) - q) > 0 \} \\ \ell(x/\theta) & \theta \in \{ \theta: (\phi(\theta) - q) \le 0 \} \end{cases}$$

3) Compute the maximum over the priors: The maximum occurs at

$$\pi^*(\theta) = \begin{cases} k\, u(\theta) & \theta \in \{ \theta: (\phi(\theta) - q) > 0 \} \\ k\, \ell(\theta) & \theta \in \{ \theta: (\phi(\theta) - q) \le 0 \} \end{cases}$$

where k is simply a normalizing constant that would make $\pi^*(\theta)$ a valid density.

4) Repeat the above steps for several values of q and numerically solve for $\overline{C}(q) = 0$.

It is obvious that we can not compute, at least in a closed form, the computational complexity of this iterative approach. The amount of computation would depend on the degree of accuracy that we wish to solve $\overline{C}(q) = 0$ equation.

Next, we will consider an approach that is based on the properties of 2-Capacities and we will use our result of Chapter 3 to directly find a closed form expression for the upper and lower posterior probabilities. See also Wasserman (1990).

5.4 Joint 2-Capacity Method

First, we will re-examine the Bayes Theorem in the context of observation and parameters, and then will proceed to discuss the joint 2-Capacity results.

5.4.1 Bayes Theorem (Revisited)

Let Xrepresent the space of measurements or observations and F be a o-field of subsets of χ, and let O be the parameter space with its corresponding σ-field B. Let $\{P(x/\theta), \theta \in O\}$ represent a family of probability measures (i.e., the sampling distributions) and Π denote the prior distribution of the parameters. We will assume that all measures have densities (with respect to some appropriate measure), and denote the sampling densities corresponding to the sarnpling distributions above as $\{f(x/\theta), \theta \in \Theta\}$ and the prior density or mass function corresponding to the prior distribution above as $\pi(\theta)$. Furthermore, Let $\chi \times \Theta$ represent the joint space of observations and parameters, and FxB be an appropriate o-algebra on this joint space. Then, Bayes theorem states that there exists a *unique* probability measure P(. x .)on $(\chi \times \Theta$, FxB), with its corresponding density $p(. , .)$, 'that has Π as its 8-marginal and $\{P(x/\theta), \theta \in O\}$ as its conditional distribution. That is,

$$P(\chi \times A) = \Pi(A) \qquad \text{for all } A \in \mathcal{B} \tag{5.4.1}$$

and for each $\theta \in O$ and any given observation x $\in X$,

$$P(\{x\} \times \Theta \mid \chi \times \{\theta\}) = f(x/\theta) \tag{5.4.2}$$

Furthermore, given an observation x, i.e., the set {x} x Θ in the joint space, we can obtain the conditional (or the posterior) probability of parameter 8, i.e., the set $\mathcal{X}$ x $\{\theta\}$, by

$$P(\mathcal{X} \times \{\theta\} | \{x\} \times \Theta) = \frac{P(\{x\} \times \{\theta\})}{P(\{x\} \times \Theta)} \tag{5.4.3}$$

The posterior density of parameter 8, given observation x, is usually denoted by $\pi(\theta/x)$ and the above expression is usually written in terms of the sampling density $f(x/\theta)$ and the prior density $\pi(\theta)$ as

$$\pi(\theta/x) = \frac{p(x,\theta)}{\int_{\Theta} f(x/\theta) d\Pi(\theta)} = \frac{f(x/\theta)\pi(\theta)}{\int_{\Theta} f(x/\theta) d\Pi(\theta)} \tag{5.4.4}$$

where above we have made use of the notation:

$$\int_A g\Pi(d\theta) = \begin{cases} \int_A g\pi(\theta) d\theta & \text{if } \theta \text{ is } \textit{continous;} \\ \sum_{\theta \in A} g\pi(\theta) & \textit{if } \theta \text{ is discrete.} \end{cases} \tag{5.4.5}$$

See DeRobertis and Hartigan (1981) for further details. The main implication of the above statements is that all we have to do is to consider the joint space of the observations and the parameters and consider the joint measure on this space. From this joint measure, we can uniquely deduce posterior related information. More specifically, given a ***set*** of priors and ***sets*** of sampling distributions, we construct the ***set*** of joint distributions. Next, we note that if the set for the joint distributions can be described by 2-Capacities, we could use our theorem of Chapter 3, to directly computed the conditional, i.e. the posterior probabilities.

5.4.2 Proposed Method Based on 2-Capacities

Let Γ^{Π} be the class of imprecise priors, $\{\Gamma_\theta^f ; \theta \in \Theta\}$ be the set of imprecise sampling densities, and let $\Gamma^{X\times\Theta}$ denote the set of resulting joint distributions. Let us assume that $\Gamma^{X\times\Theta}$ can be characterized by 2-Capacities; i.e., we can write

$$\Gamma^{X\times\Theta} = \{ P:\ P(C) \le v(C)\} \tag{5.4.6}$$

where $v(.)$ is some 2-alternating Capacity and C is a set in the product space $X \times \Theta$. Then from our Theorem in Chapter 3, eq. (3.3.6), we know that

$$v(D \mid C) = \sup_{P\in\Gamma^{X\times\Theta}} \frac{P(C\cap D)}{P(C)} \tag{5.4.7}$$

$$= \frac{v(C\cap D)}{v(C\cap D) + u(C\cap D^c)} \tag{5.4.8}$$

and similarly,

$$u(D \mid C) = \inf_{P\in\Gamma^{X\times\Theta}} \frac{P(C\cap D)}{P(C)} \tag{5.4.9}$$

$$= \frac{u(C\cap D)}{u(C\cap D) + v(C\cap D^c)} \tag{5.4.10}$$

Where typically set $D = \chi \times A$, i.e. a subset of the parameter space, and set $C = \{x_o\} \times \Theta$ is an observation in the measurement space.

Note that equations (5.4.8) and (5.4.10) provide us with a direct method to compute the conditional (i.e., posterior) upper and lower probabilities.

Example: *5.4.1* Let us reconsider Example 5.3.1 above where imprecision in both the priors and the sampling densities are described by the band models; i.e.,

$$\Gamma^{\Pi} = \{\, \pi: \ell(\theta) \le \pi(\theta) \le u(\theta) \,\} \tag{5.4.11}$$

and

$$\Gamma_{\theta}^{f} = \{\, f(./\theta):\ \ell(x/\theta) \le f(x/\theta) \le u(x/\theta) \,\}. \tag{5.4.12}$$

Then the corresponding joint space will be

$$\Gamma_{band}^{\mathcal{X}\times\Theta} = \{\, p:\ \ell(x/\theta)\,\ell(\theta) \le f(x/\theta)\,\pi(\theta) \le u(x/\theta)\,u(\theta) \,\} \tag{5.4.13}$$

$$= \{\, p:\ \ell(x;\theta) \le p(x;\theta) \le u(x;\theta) \,\} \tag{5.4.14}$$

which is also a band model.

Although the band model classes are very useful, they have two disadvantages: 1) They are usually too large and can lead to posterior ranges that are too wide and non-informative; 2) At this point, we are not aware of the 2-capacity that can characterize this class. For these two reasons, we consider the density bounded *subset* of this class. Recall that density bounded class corresponding to a band model class contains elements that are bounded by the same upper and lower bounds, are valid densities, and *do not* need renormalization. The density bounded class corresponding to (5.4.14) above is

$$\Gamma_{density\ bounded}^{\mathcal{X}\times\Theta} = \left\{\, p:\ \ell(x;\theta) \le p(x;\theta) \le u(x;\theta)\,;\ \int_{\mathcal{X}}\int_{\Theta} p(x;\theta)\,d\theta\,dx = 1 \,\right\} \tag{5.4.15}$$

or in terms of distributions

$$\Gamma_{density\ bounded}^{\mathcal{X}\times\Theta} = \{\, P:\ L(A_x \times B_\theta) \le P(A_x \times B_\theta) \le U(A_x \times B_\theta)\,;\ P(\mathcal{X}\times\Theta) = 1 \,\}$$

(5.4.16)

where $A_x \times B_\theta \in \mathcal{F} \times \beta$, and typically $A_x = \{x_o\}$ is a single observation and B_θ is a subset of parameters of interest.

We know from Chapter 2 that density bounded classes can be characterized in terrns of 2-Capacities; i.e., eq. (5.4.16) can be rewritten as

$$, \qquad = \{ \mathbf{P}: \; P(A_x \times B_\theta) \leq v(A_x \times B_\theta) \} \tag{5.4.17}$$

where

$$v(A_x \times B_\theta) = \min \{ \; U(A_x \times B_\theta), \; 1 - L((A_x \times B_\theta)^c) \} \tag{5.4.18}$$

which can be used in eq. (5.4.8) to compute the upper posterior probabilities. Lower posterior probabilities can be computed similarly.

At this time, we do not know what other classes of imprecise priors and sarnpling distributions will give rise to joint spaces that are characterized with 2-Capacities. More study is needed in this area.

It is obvious that this direct method has the lowest computational complexity of all the methods we have considered and has basically the same computational cost as the point-valued precise probabilities.

Also, this method can be extended to multiple sources under Conditional Independence (CI) assumption, as long as the resulting joint space can be charactrized in terms of 2-Capacities. Even when the joint space is not directly characterizable in terms of 2-Capacities, one can often slightly enlargen or recluce the joint space to get a new joint space which is characterizable with 2-Capacities.

Again, the only requirement for this method is that the joint space must be characterizable in terms of some joint 2-Capacity.

CHAPTER 6

INFERENCE AND DECISION-MAKING WITH IMPRECISE POSTERIOR PROBABILITIES

6.1 Introduction

Regardless of the method used to model imprecise prior probabilities and the conditional probabilities, and how they are combined to obtain posterior probabilities, the next issue is how does one proceed with these imprecise posteriors to make inferences and decisions.

In statistical inference the goal is not to make an immediate decision, but instead to provide a "summary" of the statistical evidence which a wide variety of future "users" of this evidence can easily incorporate into their own decision-making process. Posterior probabilities carry the required information. So, as far as the statistical inference is concerned, once the posterior probabilities are obtained, the task is completed.

In a decision-making process, however, given an observation, prior information and the models (or the conditional densities), rationality dictates that an action a, from the set of possible actions A, should be chosen that has minimum expected loss (or risk). See Berger (1985).

To be more specific, let Θ be the parameter space, let A be the set of all possible actions, and let λ denote the loss function; i.e.,

$$\lambda : \Theta \times \mathcal{A} \rightarrow \Re \tag{6.1.1}$$

where $\Re$ is the set of real numbers and $\lambda(\theta, a)$ is the loss incurred when action a is selected and the parameter is θ. Note that in many applications (e.g., estimation problems) $\mathcal{A} = \Theta$.

Then, the expected loss is simply

$$E[\lambda] = \frac{\int_{\Theta} \lambda(\theta,a)\, f(x/\theta)\, \Pi(d\theta)}{\int_{\Theta} f(x/\theta)\, \Pi(d\theta)} \tag{6.1.2}$$

or in terms of posterior probability $\pi(\theta \mid x)$

$$E[\lambda] = \int_{\Theta} \lambda(\theta,a)\, \pi(\theta \mid x)\, d\theta. \tag{6.1.3}$$

6.2 Upper and Lower Expected Losses

Of course, imprecise priors and imprecise sampling distributions give rise to imprecise posteriors. Let us denote the set of imprecise posteriors as $\Gamma^{\pi(./\theta)}$. Then the corresponding upper and lower posteriors can be defined, respectively, as

$$\overline{E}[\lambda] \overset{\Delta}{=} \overline{\lambda}(a) \overset{\Delta}{=} \sup_{\pi(./\theta) \in \Gamma^{\pi(./\theta)}} \int_{\Theta} \lambda(\theta,a)\, \pi(\theta \mid x)\, d\theta \tag{6.2.1}$$

and

$$\underline{E}[\lambda] \overset{\Delta}{=} \underline{\lambda}(a) \overset{\Delta}{=} \inf_{\pi(./\theta) \in \Gamma^{\pi(./\theta)}} \int_{\Theta} \lambda(\theta,a)\, \pi(\theta \mid x)\, d\theta \tag{6.2.2}$$

Note that the upper and the lower expectations are linear functions of the posteriors probabilities $\pi(\theta \mid x)$, and if the set $\Gamma^{\pi(./\theta)}$ is convex, then their

cornputation is relatively simple. In fact, if the set of the imprecise posterior probabilities $\Gamma^{\pi(./\theta)}$ can be characterized by 2-Capacities, then computation of upper and lower expected losses can be even further simplified as the following example illustrates.

Example 6.2.1: Let us assume that the set of imprecise posteriors $\Gamma^{\pi(./\theta)}$ is given by

$$\Gamma_{\varepsilon c}^{\pi(./\theta)} = \{\, \pi(.|x):\ \pi(\theta|x) = (1-\varepsilon)\,\pi_o(\theta|x) + \varepsilon\, q(\theta|x) \,\} \tag{6.2.3}$$

which is an ε - contamination model (see Chapter 2 for the definition). It is easy to see that

$$\overline{\lambda}(a) = (1-\varepsilon)\int_{\Theta} \lambda(\theta,a)\,\pi_o(\theta|x)\,d\theta + \varepsilon\,\lambda^* \tag{6.2.4}$$

where

$$\lambda^* = \sup_{\theta\in\Theta} \lambda(\theta,a)$$

For a "0-1" loss function, i.e.,

$$\lambda(\theta,a) = \begin{cases} 1 & a = \theta; \\ 0 & a \neq \theta. \end{cases} \tag{6.2.5}$$

which is a typical loss function, $\lambda^* = 1$.

Similar results can be shown for other 2-Capacity classes such as the density bounded model, etc.

There are other indirect methods for computing the upper and lower expectations which can be found in Dempster (1968), Wolfenson and Fine (1982), and Kim (1990).

6.3 Decision-Making with the Upper and Lower Expected Losses

With the usual point-value probabilities, expected losses are also point-valued and an action or a decision is made that has the minimum expected loss (or risk). For upper and lower expected losses, however, the problem is somewhat more complicated.

Let us assume that the set of actions or decisions A is finite. Then, when the upper and lower expected loss intervals are non-intersecting, the choice of an action is easy. That is, we order acts by dominance: $a_1 \succ a_2$ (read a_1 is *preferred* to a_2) if and only if

$$\underline{\lambda}(a_1) > \overline{\lambda}(a_2). \tag{6.3.1}$$

And for more than two actions, we choose the action a_i^* such that

$$a_i^* = arg\left\{ \max_j \ \underline{\lambda}(a_j) \right\} \tag{6.3.2}$$

When the upper and lower expected loss intervals overlap, however, we face the problem of indecisiveness.

When $\&(a_i) > \underline{\lambda}(a_i)$ and $\overline{\lambda}(a_j) < \overline{\lambda}(a_i)$, i.e. $[\&(a_i), \overline{\lambda}(a_j)] \subset [\underline{\lambda}(a_i), \%(a_i)]$ the intervals are nested, and it is not clear which action should be preferred and why.

What can be done, however, is to eliminate from the set of possible actions, those actions that are not preferable. That is, suppose for a_k, $k \neq i$, $k \neq j$

$$\overline{\lambda}(a_k) > \underline{\lambda}(a_i)$$

and

$$\overline{\lambda}(a_k) < \underline{\lambda}(a_j)$$

Then we eliminate a_j from further considerations and try to resolve the remaining indecision between a_i and a_k. Note also that one may face indecisiveness between a_i and a_j when,

$$\underline{\lambda}(a_j) > \underline{\lambda}(a_i)$$

and

$$\overline{\lambda}(a_j) > \overline{\lambda}(a_i)$$

There are two possibilities at this point: 1) Claim indecisiveness and require more information (e.g., in the form of more sample data for the frequentist approach), 2) Use some *ad hoc* but "reasonable" approach to resolve the problem. Let us show the above situation graphically (see Figure 6.1 below).

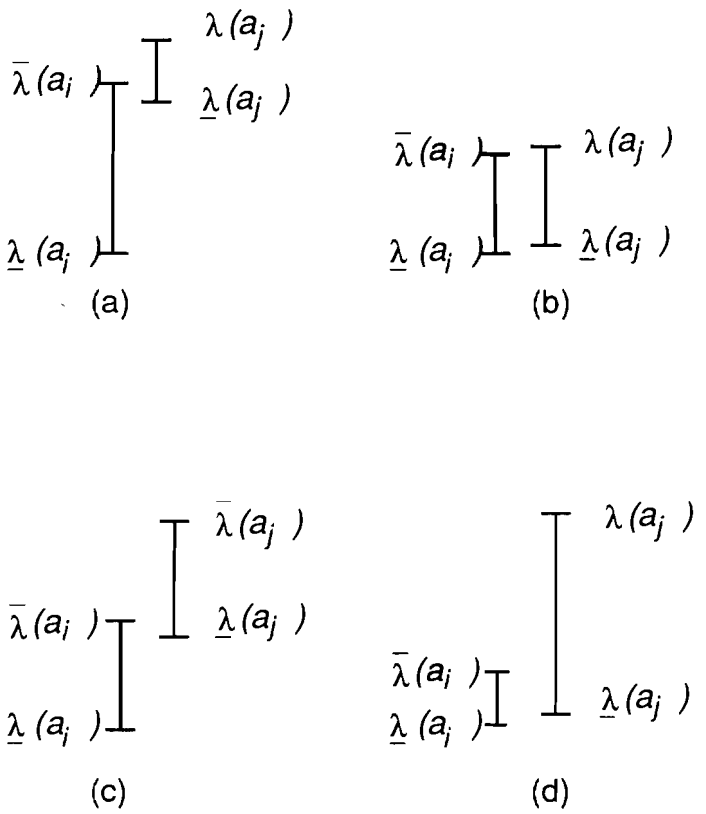

Figure 6.1 - Four Possibilities for Actions a_i and a_j with Overlapping Expected Losses : (a) $\underline{\lambda}(a_j)$ Much larger than $\underline{\lambda}(a_i)$ and $\overline{\lambda}(a_j)$ Slightly Larger than $\overline{\lambda}(a_i)$; (b) , (c) , (d) etc.

For the above scenario the following is recommended:

For case a): $a_j \succ a,$; that is a_j is preferred over a_i

For case b): a, and $a,$ are about equally preferable; this situation can happen in the point-valued expected loss problems too when the expected loss of two actions are equal. We say that we are *indifferent* about a_i and a_j and use a "tie-breaking" rule to decide.

For case c): $a, \succ a,$

For case d): $a_i \succ a_j$.

Of course, other ***ad hoc*** rules such as making decisions based on the mid-values of each interval can also be used. The main conclusion in these cases is that there is not information to make a clear decision and we need to gather more data or information. See also Loui (1986).

CHAPTER 7

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 Conclusions

The problem of representing imprecise information and combination of imprecise and uncertain information are important problems with many implications in engineering and science. The validity of any inference and decision-making mechanisim depends on the assumptions and inputs put into that system. Therefor it is essential that one models the available information carefully without either making too many unrealistic assumptions that are typically difficult to justify, or throwing away valuable information, albeit not very complete or precise, for the sake of simplicity of computation.

The goal of this thesis has been: 1) to provide realistic and useful mechanisms for describing imprecise information; 2) To come up with useful rules for combining imprecise information; 3) and finally making suggestions regarding decision-making with imprecise posteriors.

In Chapter 2, we examined various useful and natural approaches for describing imprecise information. We noted that many useful situations can easily and conveniently be characterizied in terms of Capacities. In particular, we noted that Dempster-Shafer modeling of imprecise information also is in terms of Capacities.

In Chapter 3, we drived Bayes' theorem (conditiong) for Capacities. Comparing with the results provided with the Denispter-Shafer Conditioning, which is based on Demspter-Shafer rule of combination of information, we provided another proof for inconsistency of that Demspter-Shafer rule.

Chapter 4 focused on combination of information sources. We examined some of the properties that a reasonable rule of combination of information should poses. We also showed the implications of various types of "independence" assumptions. The main conclusion here was that Bayes' combination rule with the Conditional Independence assumption had many desirable properties and avoided some of the criticism of other rules such as the Dempster-Shafer rule of combination.

In Chapter 5, we addressed the problem of combination of imprecise priors and imprecise sampling distributions. We suggested three approaches: 1) Extreme-point representation; 2) Linearization method; 3) and a direct method based on joint Capacities. We also considered the computational complexity of each approach.

In the Extreme-point approach, the available imprecise information was modeled as convex sets with identifiable extreme points. We used the extreme points of the convex imprecise priors and sampling distributions to construct the extreme points of the imprecise posterior probabilities.

In the linearization approach, we used a theorem of Lavine (Lavine 1991) to convert a nonlinear optimization problem into a set of linear optimizations. This is a powerful iterative approach.

In the direct approach, we used our theorem of Chapter 3. We noted that when the space of joint measurements and parameters can be characterized in terms of 2-Capacities, we could use the conditioning rule of Chapter 3 and directly obtain the posterior or posterior related quantities. This method, being a direct approach, has the lowest computational complexity.

Chapter 6 addressed the problem of decision-making with upper and lower expected losses. Here, we also found that with imprecise information there would be moments of indecision where a unique action or decision may not be available. We suggested a few ad hoc rules to resolve the indecisions in those situations. The main conclusion in such cases is that we simply need to gather more data.

7.2 Suggestions for Further Research

Capacities seem to be a very natural and useful tool in describing imprecise information and deserve a further examination. Bayes rule for conditioning provided in Chapter 3 is a very useful and computationally simple rule to compute the upper and lower posteriors. This rule, however, requires that the joint space of measurements and parameters be characterized in terms of 2-Capacities. Although, one can start directly with joint space s and model the imprecision in terms of 2-Capacities, this does not seem a very natural approach to us. Furthermore, even though priors and sampling distributions can easily and naturally be described in terms of 2-Capacities, at this point we do not know what family of imprecise priors and imprecise distributions would lead to joint spaces that can be characterized in terms of 2-Capacities. Considering the low cost of computational complexity of the Capacity approach, this may be a very useful direction to pursue and needs further study.

APPENDICES

Appendix A.I

Theorem:

Let $\mathcal{P}$ be a family of probability measures majorized with a 2-alternating capacity v; i.e.,

$$\mathcal{P} = \{pr: pr(A) \leq v(A)\} \tag{A.1.1}$$

and let

$$v(A/B) = \sup_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B)} \tag{A.1.2}$$

and

$$u(A/B) = \inf_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B)} \tag{A.1.3}$$

denote the upper and lower conditional probabilities over the family $\mathcal{P}$, respectively. Then, $v(A/B)$ and $u(A/B)$ can be expressed in terms of the original unconditional 2-alternating capacity v and its conjugate u as

$$v(A/B) = \frac{v(A \cap B)}{v(B \cap A) + u(B \cap A'')}$$

and

$$u(A/B) = \frac{u(A \cap B)}{u(B \cap A) + v(B \cap A'')}$$

Proof:

We give the proof for the upper conditional capacity $v(A/B)$. The proof for the lower conditional capacity is similar. First, we use fact 1) (see section 3.3) to rewrite the upper conditional capacity $v(A/B)$ as

$$v(A/B) = \sup_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B)} \tag{A.1.4}$$

$$= \sup_{pr \in \mathcal{P}} \frac{pr(A \cap B)}{pr(B \cap A) + pr(B \cap \overline{A^c})} \tag{A.1.5}$$

Now, we claim that to maximize this ratio, we can find a probability measure in $\mathcal{P}$ that *simultaneously* maximizes the numerator and minimizes the denominator. That is,

$$v(A/B) = \frac{\sup_{pr \in \mathcal{P}} pr(A \cap B)}{\sup_{pr \in \mathcal{P}} pr(B \cap A) + \inf_{pr \in \mathcal{P}} pr(B \cap A^c)} \tag{A.1.6}$$

Or using facts 2) and 3)

$$v(A/B) = \frac{\sup_{pr \in \mathcal{P}} pr(A \cap B)}{\sup_{pr \in \mathcal{P}} pr(B \cap A) + 1 + \sup_{pr \in \mathcal{P}} pr(A \cup B^c)} \tag{A.1.7}$$

This is true because

$$(B \cap A) \subset (A \cup B^c)$$

and because of the lemma stated in section 3.3. Rewriting the results in terms of the unconditional capacities v and u gives the final desired form. Q.E.D.

Appendix A.2

Theorem:

Let $\Gamma^{\ell,u}_{priors}$ and $\left\{\Gamma^{\ell,u}_{\theta_i}; i = 1 \quad, n\right\}$ represent the set of imprecise priors and the imprecise sampling distributions, respectively, where

$$\Gamma^{\ell,u}_{priors} = \left\{ \begin{matrix} \pi: \ell(\theta_i) \le \pi(\theta_i) \le u(\theta_i)\ ; i = 1, \ldots, n \\ u(\theta_i) = 1 - \sum_{k \ne i} \ell(\theta_k) \end{matrix} \right\} \qquad \text{(A.2.1)}$$

and

$$\Gamma^{\ell,u}_{\theta_i} = \left\{ \begin{matrix} f(x_j/\theta_i): \ell(x_j/\theta_i) \le f(x_j/\theta_i) \le u(x_j/\theta_i)\ ;\ j = 1, \ldots, M \\ u(x_j/\theta_i) = 1 - \sum_{k \ne j} \ell(x_k/\theta_i) \end{matrix} \right\}, \quad i = 1, \ldots, n. \qquad \text{(A.2.2)}$$

Then given an observation $x = x_j$, the posterior lower bound is given as

$$\ell(\theta_i/x_j) \overset{A}{=} \inf_{\substack{\pi \in \Gamma^{\ell,u}_{priors} \\ f(./\theta_i) \in \Gamma^{\ell,u}_{\theta_i} \\ i=1,\ldots,n}} \frac{f(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i)} \qquad \text{(A.2.3)}$$

$$= \frac{\ell(x_j/\theta_i)\,\ell(\theta_i)}{\ell(x_j/\theta_i)\,\ell(\theta_i) + u(x_j/\theta^*)\,u(\theta^*) + \sum_{\substack{k=1 \\ k \ne i,* }}^{n} u(x_j/\theta_k)\,\ell(\theta_k)} \qquad \text{(A.2.4)}$$

where $\theta^* = arg \left\{ \max_{\substack{k \\ k \ne i}} u(x_j/\theta_k) \right\}$, *provided in (A.2.4) we are not dividing by zero.*

The upper posterior probabilities can be computed as

$$u(\theta_i/x_j) = \frac{u(x_j/\theta_i)\,u(\theta_i)}{u(x_j/\theta_i)\,u(\theta_i) + \sum_{\substack{k=1 \\ k\neq i}}^{n} \ell(x_j/\theta_k)\,\ell(\theta_k)} \qquad \text{(A.2.5)}$$

Proof:

Note that

$$\ell(\theta_i/x_j) \overset{\Delta}{=} \inf_{\substack{\pi\in\Gamma^{l,u}_{priors} \\ f(./\theta_i)\in\Gamma^{l,u}_{\theta_i} \\ i=1,\ldots,n}} \frac{f(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i)} \qquad \text{(A.2.6)}$$

$$= \inf_{\pi\in\Gamma^{l,u}_{priors}} \inf_{\substack{f(./\theta_i)\in\Gamma^{l,u}_{\theta_i} \\ i=1,\ldots,n}} \frac{f(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i)}$$

$$= \inf_{\pi\in\Gamma^{l,u}_{priors}} \inf_{f(./\theta_1)\in\Gamma^{l,u}_{\theta_1}} \cdots \inf_{f(./\theta_i)\in\Gamma^{l,u}_{\theta_i}} \cdots \inf_{f(./\theta_n)\in\Gamma^{l,u}_{\theta_n}} \frac{f(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i)}$$

Note also that

$$\inf_{f(./\theta_k)\in\Gamma^{l,u}_{\theta_k}} \frac{f(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i)} = \begin{cases} \dfrac{\ell(x_j/\theta_i)\,\pi(\theta_i)}{\ell(x_j/\theta_i)\,\pi(\theta_i) + \sum_{\substack{l=1, \\ l\neq i}}^{n} f(x_j/\theta_l)\,\pi(\theta_l)} & \text{for } k = i \\ \dfrac{f(x_j/\theta_i)\,\pi(\theta_i)}{u(x_j/\theta_k)\,\pi(\theta_k) + \sum_{l=1,\,l\neq k}^{n} f(x_j/\theta_l)\,\pi(\theta_l)} & \text{for } k \neq i \end{cases}$$

thus,

$$\inf_{f(./\theta_1)\in\Gamma^{l,u}_{\theta_k}} \cdots \inf_{f(./\theta_i)\in\Gamma^{l,u}_{\theta_i}} \cdots \inf_{f(./\theta_n)\in\Gamma^{l,u}_{\theta_n}} \frac{f(x_j/\theta_i)\,\pi(\theta_i) \quad - \quad \ell(x_j/\theta_i)\,\pi(\theta_i)}{\sum_{i=1}^{n} f(x_j/\theta_i)\,\pi(\theta_i) \quad \ell(x_j/\theta_i)\,\pi(\theta_i) + \sum_{l=1,\,l\neq i}^{n} u(x_j/\theta_l)\,\pi(\theta_l)}$$

Now, we need to minimize the above quantity with respect to the priors; i.e.,

$$\ell(\theta_i/x_j) = \inf_{\pi \in \Gamma^{\mathcal{M}}_{priors}} \frac{\ell(x_j/\theta_i)\,\pi(\theta_i)}{\ell(x_j/\theta_i)\,\pi(\theta_i) + \sum_{l=1,\,l\neq i}^{n} u(x_j/\theta_l)\,\pi(\theta_l)} \qquad \text{(A.2.7)}$$

In the eq. (A.2.7) above, quantities $\ell(x_j/\theta_i)$ and $u(x_j/\theta_l)$ are nonnegative real numbers (constants) that are independent of the minimizing condition. To simplify the notation, let

$$\begin{aligned}
&\ell(x_j/\theta_i) \stackrel{\Delta}{=} c_i \\
&u(x_j/\theta_l) \stackrel{\Delta}{=} c_l \\
&\pi(\theta_i) \stackrel{\Delta}{=} z_i \\
&\ell(\theta_i) \stackrel{\Delta}{=} \ell_i \\
&u(\theta_i) \stackrel{\Delta}{=} u_i \\
&u_i - \ell_i \stackrel{\Delta}{=} d \\
&Q \stackrel{\Delta}{=} \frac{c_i\, z_i}{c_i\, z_i + \sum_{l\neq i} c_l\, z_l}
\end{aligned} \qquad \text{(A.2.8)}$$

or

$$\ell(\theta_i/x_j) = \inf\ Q \qquad \text{(A.2.9)}$$

subject to the conditions that

$$\begin{aligned}
&\ell_i \le z_i \le u_i = \ell_i + d \\
&\sum_{i=1}^{n} \ell_i + d = 1 \\
&\sum z_i = 1
\end{aligned} \qquad \text{(A.2.10)}$$

Rewriting Q as

$$Q = \frac{1}{1 + \frac{\sum_{l \neq i} c_l z_l}{c_i z_i}} \tag{A.2.11}$$

Q is minimized if and only if

$$Q' = \frac{\sum_{l \neq i} c_l z_l}{c_i z_i} \tag{A.2.12}$$

is maximized. Furthermore, since the of numerator of Q' does not contain i, Q' is maximized when $\sum_{l \neq i} c_l z_l$ is maximized and $c_i z_i$ is minimized; i.e., when Q is minimum, we have

$$z_i = \ell_i. \tag{A.2.13}$$

Note that the maximum of $\sum_{l \neq i} c_l z_l$, which is a linear combination of $z_i's$, subject to the earlier constraints which constitutes a convex set, occurs at one of the vertices of the constraint set; i.e., at

$$z_l = \begin{cases} \ell_l & l \neq j_o, i \\ \ell_{j_o} + d & l \neq i \end{cases} \tag{A.2.14}$$

so

$$\sum_{l \neq i} c_l z_l = \sum_{lei, j_o} c_l z_l + c_{j_o} z_{j_o} = \sum_{l \neq i, j_o} c_l z_l + c_{j_o}\left(\ell_{j_o} + d\right)$$

$$max \sum_{l \neq i} c_l z_l = max \left\{ \sum_{l \neq i, j_o} c_l \ell_l + c_{j_o}\left(\ell_{j_o} + d\right) \right\} = \sum_{l \neq i} c_l \ell_l + \max_{j_o \neq i} \left\{ c_{j_o}\, d \right\} \tag{A.2.15}$$

where optimal j_o, j_o^*, is selected as

$$j_o^* = arg \left\{ \max_{j_o \neq i} c_{j_o} \right\}. \tag{A.2.16}$$

This completes the proof. Proof of the upper posterior probability is similar and is omitted.

Appendix A.3

The following theorem is the extension of previous theorem to multiple sources of information under he assumption of conditional independence (CI). For simplicity of notation, we only consider two sources, though results can be easily extended to more than two sources.

Theorem:

Let us denote the observations from source 1 and source 2, with discrete random variables X and Y, respectively. Furthermore, let us assume that available information regarding each source can be expressed as lower bounds on the conditionals; i.e.,

Source 1: $$\left\{ \ell_{X/\theta}(./\theta_i) \le f_{X/\theta}(./\theta_i) \ ; i = 1, \ldots, n \right\} \quad \text{(A.3.1)}$$

Source 2: $$\left\{ \ell_{Y/\theta}(./\theta_i) \le f_{Y/\theta}(./\theta_i) \ ; i = 1, \ldots, n \right\} \quad \text{(A.3.2)}$$

and the available information regarding the priors is expressed as

Priors: $$\left\{ \ell(\theta_i) \le \pi(\theta_i) \ ; i = 1, \ldots, n \right\} \quad \text{(A.3.3)}$$

Then, under the assumption of Conditional Independence (CI) of the sources, the combined lower posterior probabilities can be computed as

$$\ell(\theta_i / x_j ; y_k) = \frac{N}{D_1 + D_2 + D_3} \quad \text{(A.3.4)}$$

where

$$N = \ell_{X/\theta}(x_j/\theta_i)\, \ell_{Y/\theta}(y_k/\theta_i)\, \ell(\theta_i)$$
$$D_1 = N = \ell_{X/\theta}(x_j/\theta_i)\, \ell_{Y/\theta}(y_k/\theta_i)\, \ell(\theta_i)$$
$$D_2 = u_{X/\theta}(x_j/\theta_*)\, u_{Y/\theta}(y_k/\theta_*)\, u(\theta_*)$$

$$D_3 = \sum_{\substack{l=1 \\ l \neq i,^*}}^{n} u_{X/\theta}(x_j/\theta_l)\, u_{Y/\theta}(y_k/\theta_l)\, \ell(\theta_k)$$

and

$$\theta^* = arg \max_{\substack{i \\ i \neq l}} \left\{ u_{X/\theta}(x_j/\theta_i)\, u_{Y/\theta}(y_k/\theta_i) \right\}. \qquad \text{(A.3.5)}$$

Proof:

Follows from previous theorem; just let

$$f_{X,Y/\theta}(x_j, y_k/\theta_i) = f_{X/\theta}(x_j/\theta_i) f_{Y/\theta}(y_k/\theta_i),$$
$$\ell_{X,Y/\theta}(x_j, y_k/\theta_i) = \ell_{X/\theta}(x_j/\theta_i) \ell_{Y/\theta}(y_k/\theta_i),$$
$$u_{X,Y/\theta}(x_j, y_k/\theta_i) = u_{X/\theta}(x_j/\theta_i) u_{Y/\theta}(y_k/\theta_i),$$

and

$$\theta^* = arg \max_{\substack{i \\ i \neq l}} \left\{ u_{X,Y/\theta}(x_j, y_k/\theta_i) \right\} = arg \max_{\substack{i \\ i \neq l}} \left\{ u_{X/\theta}(x_j/\theta_i)\, u_{Y/\theta}(y_k/\theta_i) \right\}$$

LIST OF REFERENCES

Alsina, C., E. Trillas adn L. Valverde, "On Some logical connectives for fuzzy set theory," J. *Math. Anal. appl.* 93, 15-26 (1983).

Artstein, Z., "Set-valued measures," *Trans. Amer. Math. Soc.* 165, 103-125 (1972).

Balinski, M. L., "An algorithm for finding all vertices of convex polyhedral sets", J. *Soc. Indust. Appl. Math.,* Vol. 9, No. 1, 72-88 (1961).

Banon, G., "Distinction between several subsets of fuzzy measures", *Fuzzy Sets and Systems* 5, 291-305 (1981).

Benediktsson, J. A. and P. H. Swain, "A method of statistical multisource classification with a mechanism to weight the influence of the data sources", *Proc. 'IGARSS,* 517-520 (1989).

Benediktsson, J. A., P. H. Swain and 0. K. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data", *Proc. IGARSS,* 489-492 (1989).

Berger, J. O., *Statistical decision theory and Bayesian analysis,* Springer-Verlag, Second edition (1985).

Berger, J. O., "Robust Bayesian analysis: Sensitivity to the prior," J. *Statist. Planning and Inference,* Vol. 25, 303-328 (1990).

Boole, G., "An investigation of the laws of thought," (1854); Reprinted by Dover (1958).

Buxton, R., "Modelling uncertainty in expert systems", *Int. J. Man - Machine Studies,* Vol. 31, 415-476 (1989).

Cheng Y. and R. L. Kashyap, "An axiomatic approach for combining evidence from a variety of sources," *J. Intell. and Robotic Systems* 1, 17-33 (1988).

Cheng Y. and R. L. Kashyap, "A study of associative evidential reasoning," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 11, No. 6, 623-631 (1989).

Chen, Y. and R. L. Kashyap, "Irrelevancy of evidence caused by independence assumptions", School of Electrical Engineering, Purdue University, TR-EE 86-17 (1986).

Choquet, G., "Theory of capacities," *Ann. Inst. Fourier* 5, 131-295 (1953).

Cox, R. T., "Probability, frequency and reasonable expectations," *Amer. J. Physics* 14, No. 1, 1-11 (1946).

Dempster, A., "A generalization of Bayesian inference (with discussion)," *J. Royal Statist. Soc. B* 30, 205-245 (1968).

Dempster, A., "New methods for reasoning towards posterior distributions based on sample data," *Ann. Math. Statist.* 37, 355-374 (1966).

Dempster, A., "Upper and lower probabilities induced by a multivalued mapping," *Ann. of Math. Statist.,* 38, 325-329 (1967).

DeRobertis, L. and J. A. Hartigan, "Bayesian inference using intervals of measures," *Ann. Statist.* vol. 9, No. 2, 235-244 (1981).

Domotor, Z., "Higher order probabilities," *Philosophical Studies* 40, 31-46 (1981).

Fishburn, P. C., "Analysis of decisions with incomplete knowledge of probabilities," *Op. Res.* 13 217-237 (1965).

Florens, J. P. et. al. (eds.), "Specifying statistical models," Lecture Notes in Statistics # 16, Springer-Verlag (1981).

Fryback, D., "Bayes' theorem and conditional nonindependence of data in medical diagnosis", *Computers and Biomedical Research* 11, 423-434 (1978),

Geraniotis, E., "Robust matched filters for noise uncertainty within two alternating capacity classes", *IEEE Trans. on Information Theory, VOI* 36, No. 2, 426-426 (1990).

Glymour, C., "Independence assumptions and Bayesian updating", *Artificial Intelligence* 25, 95-99 (1985).

Goicoechea, A., "Expert system for inference with imperfect knowledge: a comparative study", *J. Statistical Planning and Inference* 20, 245-277 (1988).

Hajek, P., "Combining functions for certainty degrees in consulting systems", *Int. J. Man-Machine Studies,* Vol. 22, 59-76 (1985).

Ho, Y.-C. and R. L. Kashyap, "An algorithm for linear inequalities and its applications", *IEEE Trans. on Electronic Computers,* Vol. EC-14, No. 5, 683-688 (1965).

Hoffbech, J. P. and D. A. Landgrebe, "Classification of high dimensional multispectral image data," *Fourth Annual JPL Airborne Geoscience Workshop, Arlington, Virginia,* 25-29 (1993).

Huber, P. and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," Ann. Statist., 1, 251-263 (1973a) .

Huber, P., "The use of Choquet capacities in statistics," Bull. *of* the Internat. Statist. *Inst.* Vol . XLV, Book 4, 181-188 (1973b) .

Huber, P., Robust Statistics, Wiley, 260-263 (1981).

Jaynes, E. T., "Prior probabilities," *IEEE* Trans. Syst. Sci. Cybern., vol. SSC-4, 227-241 (1968).

Johnson, R. W., "Independence and Bayesian updating methods", Artificial Intelligence 29, 217-222 (1986).

Karmarkar, N., "A new polynomial-time algorithm for linear programming", "*Combinatorica* 4 (4), 373-395 (1984).

Kassam, S. A., "Robust hypothesis testing for bounded classes of probability densities," *IEEE* Trans. Inform. Theory, Vol. IT-27, No. 2, 242-247 (1981).

Kennes, R., "Computational aspects of the Mobius transformation of graphs",*IEEE* Trans on Systems, Man, and Cybernetics, Vol. 22, No. 2, 201-223 (1992).

Kim, B. and D. A. Landgrebe, "Hierarchical classifier design in high dimensional, numerous class cases," IEEE Trans on *Geoscience* and Remote Sensing, Vol. 29, No. 4, 518-528 (1991).

Kim, H., "A method of classification for multisource data in remote sensing based on interval-valued probabilities", Ph.D. thesis, School of Electrical Engineering, Purdue University, 1990.

Klir, G. J. and T.A. Folger, "Fuzzy sets, uncertainty, and information," Prentice-Hall (1988).

Koopman, B. O., "The axioms and algebra of intuitive probability"' Ann. Math. 41 269-278 (1940).

Kyburg, H. E., Jr., "Bayesian and non-Bayesian evidential updating," *Artificial* Intelligence 271-293 (1987).

Landgrebe, D. A., "A perspective on the analysis of hyperspectral data," Proceedings of International Geoscience and *Remote* Sensing Symposium *(IGARSS'93),* Tokyo, 1362-4 (1993).

Lee, C. and D. A. Landgrebe, "Decision boundary feature extraction for non-parametric classification," *IEEE* Trans. on Systems, Man, and Cybernetics, Vol. 31, No. 4, 792-800 (1993).

Lee, T., J. A. Richards and P. H. Swain, "Probabilistic and evidential approaches to multisource data analysis", *IEEE* Trans. Geos. and Remote Sensing, Vol. GE-25, 283-293 (1987).

Lemmer, J. F., "General Bayesian updating of incompletely specified distributions", Large Scale Systems 5, 51-68 (1983).

Lavine, M., "Sensitivity in Bayesian statistics: The prior and the likelihood," J. Arner. Statist. Assoc. Vol. 86, 396-399 (1991).

Loui, R. P., "Decisions with indeterminate probabilities", *Theory* and Decision 21, 283-309 (1986).

Matheiss, T. M. and D. S. Rubin, "A survey and comparison of methods for findingall vertices of convex polyhedral sets", Mathematics of Operations Research, Vol 5, No. 2, 167-185 (1980).

Menger, K., Statistical metrics, Proc. Nat. Acad. Sci. U.S.A. 28, 535-537 (1942).

Negoiwta, C. V. and D. Ralescu, Simulation, *Knowledge-based* Computing, and Fuzzy Statistics, Van Nostrand Reinhold (1987).

Orponen, P.,"Dempster's rule of combination is #p-complete," Artificial Intelligence 44, 245-253 (1990).

Pednault, E. P. D., S. W. Zucker and L. V. Muresan, "On the independence assumption underlying subjective bayesian updating", Artificial Intelligence16,213-222 (1981).

Potter, J. M. and B.D. Anderson, "Partial prior information and decisionmaking, " *IEEE* trans. Syst. Man Cybern., Vol. SMC-10, No.3, 125-133 (1980).

Puri, M. L. and Dan A. Ralescu, "Strong law of large numbers with respect to a set-valued probability measure," Ann. Prob., Vol. 11, No.4, 1051-1054 (1983).

Safavian, S. R. and D. A. Landgrebe, "A survey of decision tree classifier methodologies," *IEEE* Trans. on Systems, Man, and cybernetics, Vol. 21, No. 3, 660-674 (1991).

Schocken, S. and P. R. Kleindorfer, "Artificial intelligence dialects of the Bayesian belief revision language," *IEEE* Trans. Systems, Man, and Cybern. Vol. 19, No. 5, 1106-1121 (1989).

Schweizer, B. and A. Sklar, Probabilistic metric spaces, New 'York: North-Holland (1983).

Shafer, G., "A mathematical theory of evidence," Princeton Univ. Press (1976).

Shafer, G., "Belief functions and parametric models (with discussion)," J. *Roy. Statist Soc. Ser. B* 44, 322-352 (1982).

Shore, J. E. and R.M. Gray, "Minimum cross-entropy pattern classification and cluster analysis," *IEEE Trans. Patt. Anal. and Mach. Intell.*, Vol. 4, No.1, 11-17 (1981).

Smets, P., "The combination of evidence in the transferable belief model," *IEEE Trans. Patt. Anal. and Mach. Intell.*, Vol. 12, No.5, 447-458 (1990).

Smets, P., "Medical diagnosis: Fuzzy sets and degrees of belief," *Fuzzy Sets and Systems,* Vol. 5, 259-266 (1981).

Smets, P., "Belief functions versus probability functionc," In Uncertainty in Intelligent Systems; eds. B. Bouchon, L. Saitta and R. R. Yager, 17-24 (1988).

Smith, C. A. B., "Consistency in statistical inference and decision (with discussion)," J. *Roy. Statist. Soc. Ser.* B, 23, 1-25 (1961).

Stirling, W. C., "Convex Bayes decision theory," *IEEE Trans. on System, Man, and Cybernetics,* Vol. 1, No. 1, 173-183 (1991).

Sundberg, C. and C. Wagner, "Generalized finite differences and Bayesian Conditioning of Choquet capacities," *Submitted to* J. *Theory of Prob. (1994 a)*

Sundberg, C. and C. Wagner, "Characterization of momotone and two-monotone Capacities," *Submitted to* J. *Theory of Prob. (1994 b)*

Swain, P. H., J. A. Richards and T. Lee, "Multisource data analysis in remote sensing and geographic information processing", *Proc. 11th Int. Sym. on Machine Processing of Remotely Sensed Data,* 211-217 (1985).

Vastola, K. and V. Poor, "On generalized band models in robust detection and filtering," *Proc. 14th Conf. Inform. Sci. Syst., Princeton, N.J.,* 1-5 (1980).

Voorbraak, F., "On the justification of Dempster's rule of combination", *Artificial Intelligence* 48, 171-197 (1991).

Walley, P., "Belief function representations of statistical evidence," *Ann. Statist.* Vol. 15, No. 4, 1439-1465 (1987).

Walley, P. and T.L. Fine, "Toward a frequentist theory of upper and lower probability," *Ann. Statist.,* Vol. 10, No.3, 741-761 (1983).

Wasserman, L. A. and J. B. Kadane, "Bayes' theorem for choquet capacities," *Ann. Statistics,* Vol. 18, No. 3, 1328-1339 (1990).

Wasserman, L. A., M. Lavine and R. L. Wolpert, "Linearization of Bayesian robustness problems," *J. Statist. Plann. Inference,* Vol. 37, 307-316 (1993).

Weber, S., "A general concept of fuzzy connectives, negations and implications based on t-norms and t-conorms," *Fuzzy Sets and Systems* 11, 115-134 (1983).

Williams, P. M., "Indeterminate probabilities," In Formal methods in the methodology of empirical sciences, M. Przelecki, K. Szaniawski, and R. Wojeiki (eds.) Reidel (1976).

Wolfenson, M. and T.L. Fine, " Bayes-like decision making with upper and lower probabilities," *J. Amer. Statsit. Assoc.* 77, 80-88 (1982).

Zadeh, L. A." On the validity of Dempster's rule of combination," memorandum No. UCB/ERL M79/24, Univ. of Calif., Berkeley (1979) .

Zadeh, L. A.," Review of : 'A mathematical theory of evidence' by G. Shafer," *Artificial Intelligence,* 81-83 (1984) .

Zadeh, L. A.," A simple view of The Dempster-Shafer theory of evidence and its implication for the rule of combination," *Artificial Intelligence,* 85-90 (1*986)* .