Uncertainty Modelling for Intelligent Systems

Tim Lawson

October 26, 2023

These notes are based on the lecture notes for the unit Uncertainty Modelling for Intelligent Systems in 2023/24 at the University of Bristol.

Contents

Theorems

1 Possible worlds and quantitative uncertainty

2 Probability theory

2.1 Introduction

Probability theory is the best-known theory of uncertainty. Definition 2.1.1 states that probability measures are additive uncertainty measures.

Definition 2.1.1 (Probability measure)**.** *A probability measure is a function* $P : 2^W \to [0, 1]$ *such that:*

P1 $P(W) = 1$ *and* $P(\emptyset) = 0$

P2 $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Additive uncertainty measures are monotonic.

Theorem 2.1.2 (Monotonicity)**.**

$$A \subseteq B \Rightarrow P(A) \leq P(B) \tag{1}$$

Proof. Let $A, B \subseteq W$ such that $A \subseteq B$. By definition, $B = A \cup (B \cap A^c)$ and $A \cap (B \cap A^c) = \emptyset$. From definition 2.1.1, $P(B) = P(A) + P(B \cap A^c)$ and $P(B \cap A^c) \in [0, 1]$, hence $P(B) \geq P(A)$. □

Theorem 2.1.3 (Complement)**.**

$$P(A^c) = 1 - P(A) \tag{2}$$

Proof. By definition, $A \cup A^c = W$ and $A \cap A^c = \emptyset$. From definition 2.1.1, $P(A \cup A^c) = P(A) + P(A^c) = 1$, hence 2. □

Theorem 2.1.4 (General additivity)**.**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3}$$

Proof. By definition, $A = (A \cap B) \cup (A \cap B^c)$, $B = (B \cap A) \cup (B \cap A^c)$ and $A \cup B = (A \cap B) \cup (A \cap B^c) \cup (B \cap A^c)$. $A \cap B$, $A \cap B^c$, and $B \cap A^c$ do not overlap. Hence, from definition 2.1.1:

$$P(A) = P(A \cap B) + P(A \cap B^c) \tag{4}$$

$$P(B) = P(B \cap A) + P(B \cap A^c) \tag{5}$$

$$P(A \cup B) = P(A \cap B) + P(A \cap B^c) + P(B \cap A^c) \tag{6}$$

Substituting 4 and 5 into 6 gives 3. □

Theorem 2.1.5 (Theorem of total probability)**.**

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c) \tag{7}$$

Definition 2.1.6 (Probability distribution)**.** *A probability distribution is a function* $P(\{w\})$ *or* $P(w) : w \in W$*.*

A probability measure is completely determined by its associated probability distribution. Let $A \subseteq W$ be a proposition. By definition 2.1.1:

$$P(A) = P(\bigcup_{w \in A} \{w\}) = \sum_{w \in A} P(w) \tag{8}$$

It is more practical to work with probability measures than general uncertainty measures. For $n = |W| - 1$, a general uncertainty measure is defined by $2^n - 2$ values, whereas a probability measure is defined by $n - 1$ values.

2.2 Updating probabilities and Bayesian reasoning

If an agent learns that a proposition A is true, then it should update its probability distribution to reflect that $w^* \in A$. If an agent learns that a proposition A is false, then it cannot update its probability distribution.

Definition 2.2.1 (Conditional probability)**.** *Let P be a probability measure such that $P(B) > 0$. The conditional probability of A given B is:*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{9}$$

Let $H = \{H_i \mid i \in 1..k\} \subseteq W$ be a set of hypotheses. H are:

- *mutually exclusive* if $H_i \cap H_j = \emptyset \ \forall \ i \neq j$; and
- *exhaustive* if $\bigcup_{i=1}^{k} H_i = W$.

In the context of Bayes' theorem:

- $P(H_i \mid E)$ are the *posterior* probabilities;
- $P(H_i)$ are the *prior* probabilities; and
- $P(E \mid H_i)$ are the *likelihoods*.

It is generally impractical to evaluate the posterior probabilities. But it is more practical to estimate the prior probabilities and the likelihoods. Bayes' theorem can be used to estimate the posterior probabilities.

Theorem 2.2.2 (Bayes' theorem)**.** *Let $\{H_i \mid i \in 1..k\} \subseteq W$ such that $H_i \cap H_j = \emptyset \ \forall \ i \neq j$ and $\bigcup_{i=1}^{k} H_i = W$. Then, for $E \subseteq W$ such that $P(E) > 0$:*

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{\sum_{j=1}^{k} P(E \mid H_j)P(H_j)} \ \forall \ i \in 1..k \tag{10}$$

Proof. By definition 2.2.1:

$$P(E \mid H_i) = \frac{P(H_i \cap E)}{P(H_i)} \ , \ P(H_i \mid E) = \frac{P(E \cap H_i)}{P(E)} \tag{11}$$

Therefore:

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{P(E)} \tag{12}$$

$\{H_i \mid i \in 1..k\}$ is a partition of W, hence:

$$\sum_{i=1}^{k} P(H_i \mid E) = 1 \tag{13}$$

Substituting 13 into 12 gives:

$$1 = \sum_{i=1}^{k} \frac{P(E \mid H_i)P(H_i)}{P(E)} = \frac{\sum_{i=1}^{k} P(E \mid H_i)P(H_i)}{P(E)} \tag{14}$$

Substituting 14 into 13 gives 10. □

Conditional inference applies a general rule to a specific instance:

- $P(A \mid B)$ is a conditional probability defined on a set of instances; and
- $P(B)$ is a probability defined on a specific instance.

In Bayes' theorem, proposition E is certainly true, i.e., $w^* \in E$. Jeffrey's rule is a generalisation of Bayes' theorem where $P(E) \in [0, 1]$. Its result is a new probability measure P'.

Theorem 2.2.3 (Jeffrey's rule)**.**

$$P'(H) = P(H \mid E)P'(E) + P(H \mid E')(1 - P'(E)) \tag{15}$$

Theorem 2.2.3 is an extension of 2.1.5 that permits inference about specific instances from general conditional probabilities.

Bayesian reasoning strongly depends on the prior probabilities. Any posterior probability can be obtained by selecting an appropriate prior probability. I.e., the likelihoods alone do not determine the posterior probabilities.

Definition 2.2.4 (Bayes factor)**.** *Given evidence E and a hypothesis H, the Bayes factor is:*

$$F = \frac{P(E \mid H)}{P(E \mid H^c)} \tag{16}$$

2.3 Prior probabilities

Bayesian reasoning is difficult without prior knowledge of the hypotheses. The most common approach to this problem is *Laplace's principle of insufficient reasoning*:

> In the absence of any other information, all hypotheses under consideration should be assumed to be equally probable, i.e., the probability distribution should be *uniform*.

The uniform distribution is the least informative (theorem 3.1.3).

Probability theory conflates uncertainty (a lack of knowledge as to the true possible world, quantified by a probability measure) and ignorance (a lack of knowledge that makes it difficult to quantify one's beliefs). A different approach is to use a theory of uncertainty that differentiates between uncertainty and ignorance, e.g., Dempster-Shafer theory.

3 Probabilistic reasoning

Classical logic is based on propositions and rules that are known with certainty. This is generally impossible. Probabilistic reasoning is based on uncertain knowledge. There are generally multiple probability distributions that are consistent with a knowledge-base.

3.1 Probabilistic knowledge-based reasoning

One approach to probabilistic reasoning is to identify the set of probability distributions $\mathbb{P}(K)$ that are consistent with a knowledge-base K, then select a probability distribution based on some principle.

Definition 3.1.1 (Probabilistic knowledge-base)**.** *A probabilistic knowledge-base K is a set of linear equations on P:*

$$K = \left\{ \sum_{i=1}^{n_j} a_{ij} P(A_{ij}) = b_j \ : \ j = 1..m \ \wedge \ A_{ij} \subseteq W \ \wedge \ a_{ij}, b_k \in \mathbb{R} \right\} \tag{17}$$

One principle is to select the distribution that has minimal information. Entropy is a measure of the information content of a probability distribution.

Definition 3.1.2 (Entropy)**.** *Let $W = \{w_i \mid i \in 1..n\}$ and $P(w_i) = p_i : i \in 1..n$. Entropy is:*

$$H(\{p_i \mid i \in 1..n\}) = \sum_{i=1}^{n} -p_i \log_2(p_i) \tag{18}$$

If the knowledge-base is linear, then there is a single probability distribution with maximum entropy (theorem 3.1.3).

Theorem 3.1.3 (Maximum entropy distribution)**.** *Let $W = \{w_i \mid i \in 1..n\}$. The maximum entropy distribution in $\mathbb{P}$ is $P(w_i) = \frac{1}{n}$.*

Proof. Let $P(w_i) = p_i : i \in 1..n$. Without loss of generality, let $p_n = 1 - \sum_{i=1}^{n-1} p_i$ such that H is a function of $\{p_i \mid i \in 1..n-1\}$.

H is maximal when $\frac{\partial H}{\partial p_i} = 0 : i \in 1..n-1$. By the product and chain rules:

$$\frac{\partial H}{\partial p_i} = \frac{\partial}{\partial p_i}(-p_i \log_2(p_i)) - \frac{\partial}{\partial p_i}(-p_n \log_2(p_n)) \tag{19}$$

$$= -\log_2(p_i)\frac{\partial p_i}{\partial p_i} - p_i \frac{\partial \log_2(p_i)}{\partial p_i} + \log_2(p_n)\frac{\partial p_n}{\partial p_i} + p_n \frac{\partial \log_2(p_n)}{\partial p_i} \tag{20}$$

By the properties of logarithms:

$$\frac{\partial \log_2(p_i)}{\partial p_i} = \frac{1}{\ln 2}\frac{\partial \ln p_i}{\partial p_i} = \frac{1}{p_i \ln 2} \tag{21}$$

$$\frac{\partial \log_2(p_n)}{\partial p_i} = \frac{1}{\ln 2}\frac{\partial \ln p_n}{\partial p_i} = \frac{1}{\ln 2}\frac{\partial p_n}{\partial p_i}\frac{\partial \ln p_n}{p_n} = -\frac{1}{p_n \ln 2} \tag{22}$$

Substituting equations 21 and 22 into equation 20 yields:

$$\frac{\partial H}{\partial p_i} = -\log_2(p_i) - \log_2(p_n) = 0 \tag{23}$$

That is, $\log_2(p_i) = \log_2(p_n)$ and $p_i = p_n : i \in \{1..n-1\}$. Since $\sum_{i=1}^{n} p_i = 1$ and $p_n = 1 - \sum_{i=1}^{n-1} p_i$, $p_i = \frac{1}{n}$. □

Another principle is to define a uniform distribution over the set of probability distributions, then select the expected value of that distribution, i.e., its *centre of mass* (definition 3.1.4).

Definition 3.1.4 (Centre of mass)**.** *Let $\mathbb{P}(K)$ be a set of probability distributions that are consistent with a knowledge-base K. The centre of mass is:*

$$P(w_i) = \frac{\int_{\mathbb{P}(K)} p_i \, d\mathbb{P}(K)}{\int_{\mathbb{P}(K)} d\mathbb{P}(K)} \tag{24}$$

3.2 Probability logic

Probability theory is not truth-functional. Knowledge of $P(A)$ and $P(B)$ does not imply knowledge of, e.g., $P(A \cap B)$, but it does imply upper and lower bounds (theorems 3.2.1 and 3.2.2).

Theorem 3.2.1 (Intersection bounds)**.** *Let $A, B \subseteq W$.*

$$\max(0, P(A) + P(B) - 1) \leq P(A \cap B) \leq \min(P(A), P(B)) \tag{25}$$

Proof. By definition 2.1.1:

$$P(A) = P(A \cap B) + P(A \cap B^c) \tag{26}$$
$$P(B) = P(A \cap B) + P(A^c \cap B) \tag{27}$$

Since $P(A \cap B^c) \geq 0$ and $P(A^c \cap B) \geq 0$:

$$P(A \cap B) \leq \min(P(A), P(B)) \tag{28}$$

By theorem 2.1.4:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \tag{29}$$

Since $P(A \cup B) \leq 1$ and $P(A \cap B) \geq 0$:

$$P(A \cap B) \geq \max(0, P(A) + P(B) - 1) \tag{30}$$

□

Theorem 3.2.2 (Union bounds)**.** *Let $A, B \subseteq W$.*

$$\max(P(A), P(B)) \leq P(A \cup B) \leq \min(P(A) + P(B), 1) \tag{31}$$

Proof. By theorems 2.1.4 and 3.2.1:

$$P(A \cup B) \geq P(A) + P(B) - \min(P(A), P(B)) = \max(P(A), P(B)) \tag{32}$$

By theorem 3.2.1:

$$P(A \cup B) \leq P(A) + P(B) - \max(0, P(A) + P(B) - 1) \tag{33}$$
$$\leq \min(P(A) + P(B), 1) \tag{34}$$

□

Theorems 2.1.3, 3.2.1, and 3.2.2 define a truth-functional logic for probability intervals. For a proposition $A \subseteq W$, it can be used to infer upper and lower bounds on $P(A)$, i.e., $P(A) \in [L(A),\ U(A)]$.

Theorem 3.2.3 (Complement rule)**.**

$$P(A^c) \in [1 - U(A),\ 1 - L(A)] \tag{35}$$

Theorem 3.2.4 (Intersection rule)**.**

$$P(A \cap B) \in [\max(0, L(A) + L(B) - 1),\ \min(U(A), U(B))] \tag{36}$$

Theorem 3.2.5 (Union rule)**.**

$$P(A \cup B) \in [\max(L(A), L(B)),\ \min(U(A) + U(B),\ 1)] \tag{37}$$

Theorem 3.2.6 (Conditional rule)**.**

$$P(A) \in [L(A \mid B)L(B),\ (U(A \mid B) - 1)U(B) + 1] \tag{38}$$

Proof. By theorem 2.1.5 and definition 2.1.1, i.e., $P(A \mid B^c) \in [0, 1]$:

- If $P(A \mid B^c) = 0$, then $P(A) = P(A \mid B)P(B)$ (lower bound).
- If $P(A \mid B^c) = 1$, then $P(A) = (P(A \mid B) - 1)P(B) + 1$ (upper bound).

□

Theorem 3.2.7 (Jeffrey's rule)**.**

$$P'(A) \in [L(A \mid B)L(B),\ (U(A \mid B) - 1)U(B) + 1] \tag{39}$$

The upper and lower bounds of this logic are wider than probability theory constrains them to be. This is because the logic does not account for logical dependencies between propositions and their relations to definition 2.1.1.

3.3 Bayesian networks

3.3.1 Joint and marginal probability distributions

Definition 3.3.1 (Joint probability distribution)**.** *The joint probability distribution of* $X = \{X_i : W \to \Omega \mid i \in 1..n\}$ *is:*

$$\begin{aligned} P(X) &= P\Big(\bigwedge_{i=1}^{n} X_i = x_i\Big) \\ &= P(\{w : X_1(w) = x_1, \ldots, X_n(w) = x_n\}) \end{aligned} \tag{40}$$

If $|\{X_i(w) : w \in W\}| = k_i$, then the joint probability distribution has $k_i^n - 1$ values, i.e., the number of values grows exponentially with the dimension of the random variables (the *curse of dimensionality*).

Definition 3.3.2 (Marginal probability distribution)**.** *The marginal probability distribution of* $X_i : W \to \Omega$ *is:*

$$P(X_i = x_i) = P(\{w : X_i(w) = x_i\})\ \forall\ i \in 1..n \tag{41}$$

Theorem 3.3.3 (Joint and marginal probability distributions)**.** *The joint and marginal probability distributions are related by:*

$$P(X_i = x_i) = \sum_{j \neq i} \sum_{x_j} P\big(\bigwedge_{j=1}^{n} X_j = x_j\big) \ \forall \ i \in 1..n \tag{42}$$

I.e., the marginal probability distribution of X_i is the sum of the joint probability distributions of X over all values x_j where $j \neq i$.

3.3.2 Independence

By definition 3.3.4, if X are *independent*, then the joint probability distribution of X is defined by its *marginal* distributions. Then, the number of values grows only linearly with the dimension of the random variables.

Definition 3.3.4 (Independence)**.** *Let $X = \{X_i \mid i \in 1..j\}$ be a set of random variables. X are independent if:*

$$P\big(\bigwedge_{i=1}^{n} X_i = x_i\big) = \prod_{i=1}^{n} P(X_i = x_i) \ \forall \ x_i \in \{X_i(w) : w \in W\} \tag{43}$$

3.3.3 Conditional distributions

We cannot generally assume that random variables are independent. But some random variables are not directly dependent on others. The formalization of this intuition helps to make probabilistic reasoning computationally feasible.

Definition 3.3.5 (Conditional distribution)**.** *Let $X = \{X_i \mid i \in 1..j\}$ be a set of random variables. Without loss of generality, let $\{X_1, X_2\} \in \mathbb{X}$ form a partition of X:*

$$X_1 = \{X_i \mid i \in 1..k-1\}$$
$$X_2 = \{X_i \mid i \in k..j\}$$

The conditional probability of $\bigwedge_{i=1}^{k-1} X_i = x_i$ given $\bigwedge_{i=k}^{j} X_i = x_i$ is:

$$P\big(\bigwedge_{i=1}^{k-1} X_i = x_i \mid \bigwedge_{i=k}^{j} X_i = x_i\big) = \frac{P(\bigwedge_{i=1}^{j} X_i = x_i)}{P(\bigwedge_{i=k}^{j} X_i = x_i)} \tag{44}$$

The denominator of equation 44 is the joint probability distribution:

$$P(\bigwedge_{i=k}^{j} X_i = x_i) = \sum_{x_1} \dots \sum_{x_{j-1}} P(\bigwedge_{l=1}^{j} X_l = x_l) \tag{45}$$

I.e., if an agent receives information that $\bigwedge_{l=k}^{j} X_l = x_l$, then to update its probabilities for the other random variables $\{X_i \mid i \in 1..k-1\}$, it evaluates the conditional probability of the other random variables given the information.

Definition 3.3.6 (Conditional independence)**.** *Let $X = \{X_i \mid i \in 1..j\}$ be a set of random variables. Without loss of generality, let $\{X_1, X_2, X_3\} \in \mathbb{X}$ form a partition of X:*

$$X_1 = \{X_i \mid i \in 1..k-1\}$$
$$X_2 = \{X_i \mid i \in k..l-1\}$$
$$X_3 = \{X_i \mid i \in l..j\}$$

Then X_1 are conditionally independent of X_3 given X_2 if:

$$P\big(\bigwedge_{i=1}^{k-1} X_i = x_i \mid \bigwedge_{i=k}^{j} X_i = x_i\big) = P\big(\bigwedge_{i=1}^{k-1} X_i = x_i \mid \bigwedge_{i=k}^{l-1} X_i = x_i\big) \;\forall\; x_i \in \{X_i(w) : w \in W\} \quad (46)$$

I.e., if an agent knows the values of $\{X_i \mid i \in k..l-1\}$, then it can ignore the values of $\{X_i \mid i \in l..j\}$ when updating its probabilities for $\{X_i \mid i \in 1..k-1\}$.

3.3.4 Bayesian networks

A Bayesian network is a graphical model of probabilistic reasoning with multiple random variables. It is a compromise between independence and dependence: it assumes *independence where possible* and *dependency where necessary.*

Definition 3.3.7 (Directed graph)**.** *A directed graph is an ordered pair (V, E) where $V = \{v_i \mid i \in 1..j\}$ is a set of vertices and E is a binary relation on V that defines a set of edges.*

Definition 3.3.8 (Directed acyclic graph)**.** *A directed graph (V, E) is acyclic if there is no sequence of vertices $v_i \dots v_k$ where $v_i = v_k$ and $(v_l, v_{l+1}) \in E$ for all $l \in 1..k-1$.*

Definition 3.3.9 (Bayesian network)**.** *A Bayesian network is:*

- *a directed acyclic graph (V, E) where each vertex $v_i \in V$ is a random variable $X_i : W \to \Omega$ and $(X_i, X_j) \in E$ only if $j < i$; and*
- *a joint probability distribution on $\{X_i \mid i \in 1..k\}$ where:*

$$P(X_i \mid \{X_j \mid j \in 1..i-1\}) = P(X_i \mid \Pi(X_i)) \quad (47)$$

and $\Pi(X_i) = \{X_k : (X_k, X_i) \in E\}$ is the set of parent vertices of X_i.

For a Bayesian network, we assume that X_i is conditionally independent of its *indirect* causes $\{X_j \mid j \in 1..i-1\} - \Pi(X_i)$ given its *direct* causes $\Pi(X_i)$. The joint probability distribution of X is determined by its conditional distributions $P(X_i \mid \Pi(X_i)) \;\forall\; i \in 1..k$ (theorem 3.3.10).

Theorem 3.3.10 (Joint probability distribution of a Bayesian network)**.**

$$P(X) = \prod_{i=1}^{k} P(X_i \mid \Pi(X_i)) \quad (48)$$

Proof. Trivially:

$$P(\{X_i \mid i \in 1..k\}) = \prod_{i=1}^{k} P(X_i \mid \{X_j \mid j \in 1..i-1\}) \tag{49}$$

By the conditional independence assumptions of definition 3.3.9:

$$\prod_{i=1}^{k} P(X_i \mid \{X_j \mid j \in 1..i-1\}) = \prod_{i=1}^{k} P(X_i \mid \Pi(X_i)) \tag{50}$$

□

3.3.5 Numbers of probability values

For n binary random variables, the number of probability values in the joint probability distribution of:

- a fully dependent model is $2^n - 1$;
- a fully independent model is $2n - 1$; and
- a Bayesian network is $\sum_{i=1}^{n} 2^{|\Pi(X_i)|}$.

Theorem 3.3.11 (Number of probability values for a Bayesian network)**.** *For n binary random variables, the joint probability distribution of a Bayesian network has between n and $2^n - 1$ values.*

Proof. The minimum of $|\Pi(X_i)|$ is 0, in which case:

$$\sum_{i=1}^{n} 2^{|\Pi(X_i)|} = \sum_{i=1}^{n} 2^0 = n \tag{51}$$

The maximum of $|\Pi(X_i)|$ is $i - 1$, in which case:

$$\sum_{i=1}^{n} 2^{|\Pi(X_i)|} = \sum_{i=1}^{n} 2^{i-1} = \sum_{i=0}^{n-1} 2^i = \frac{1(1-2^n)}{1-2} = 2^n - 1 \tag{52}$$

□

For n random variables, where the random variable X_i has k_i possible values, i.e., $|\{X_i(w) : w \in W\}| = k_i \ \forall \ i \in 1..n$, the number of values in:

- the conditional distributions $P(X_i \mid \Pi(X_i))$ is $(k_i - 1) \prod_{X_j \in \Pi(X_i)} k_j$; and
- the joint probability distribution of a Bayesian network is $\sum_{i=1}^{n} (k_i - 1) \prod_{X_j \in \Pi(X_i)} k_j$.

Generally, the number of values for a Bayesian network is inversely proportional to the numbers of direct causes of the random variables.

4 Ignorance and uncertainty

Probability theory conflates ignorance and uncertainty. This section describes approaches to explicitly modelling them.

4.1 Dempster-Shafer theory

In Dempster-Shafer theory, uncertainty is quantified by two measures:

- *belief* (evidence that implies a proposition); and
- *plausibility* (evidence that is consistent with a proposition).

Definition 4.1.1 (Belief and plausibility measures)**.** *Let W be a set of possible worlds and $A \subseteq W$ be a proposition. A mass function $m : 2^W \to [0, 1]$ generates:*

- *a belief measure* $\text{bel} : 2^W \to [0, 1]$ *such that:*

$$\text{bel}(A) = \sum_{B \subseteq W : B \subseteq A} m(B) \;\forall\; A \subseteq W \tag{53}$$

- *a plausibility measure* $\text{pl} : 2^W \to [0, 1]$ *such that:*

$$\text{pl}(A) = \sum_{B \subseteq W : B \cap A \neq \emptyset} m(B) \;\forall\; A \subseteq W \tag{54}$$

I.e., for a proposition $A \subseteq W$:

- the *belief* in A is the sum of the masses of the subsets of W that are subsets of A, i.e., of the evidence that implies A; and
- the *plausibility* of A is the sum of the masses of the subsets of W that intersect A, i.e., of the evidence that is consistent with A.

Theorem 4.1.2 (Relations of belief and plausibility measures)**.** *Let W be a set of possible worlds and $A \subseteq W$ be a proposition.*

$$\text{bel}(A) \leq \text{pl}(A) \qquad \forall\; A \subseteq W \tag{55}$$

$$\text{pl}(A) = 1 - \text{bel}\, A^c \;\forall\; A \subseteq W \tag{56}$$

Proof. $\emptyset \neq B \subseteq A \;\Rightarrow\; B \cap A \neq \emptyset$, hence equation 55.
$B \cap A \neq \emptyset \;\Leftrightarrow\; B \not\subseteq A^c$, hence equation 56:

$$\text{pl}(A) = \sum_{B \subseteq W : B \cap A \neq \emptyset} m(B) \;=\; \sum_{B \subseteq W : B \not\subseteq A^c} m(B)$$

$$= 1 - \sum_{B \subseteq W : B \subseteq A^c} m(B) \;=\; 1 - \text{bel}(A^c)$$

□

Theorem 4.1.3 (Relation to probability theory)**.** *If $m : 2^W \to [0, 1]$ is a mass function such that $\sum_{w \in W} m(\{w\}) = 1$, then:*

$$\text{bel}(A) = \text{pl}(A) = P(A) = \sum_{w \in A} m(\{w\}) \tag{57}$$

I.e., if m is non-zero only for singletons, then each piece of evidence identifies a single possible world. Generally, belief and plausibility measures do not satisfy definition 2.1.1.

Theorem 4.1.4 (Belief measures are super-additive)**.**

$$\text{bel}(A \cup B) \geq \text{bel}(A) + \text{bel}(B) \tag{58}$$

Proof. Let W be a set of possible worlds and $A, B \subseteq W$ be propositions such that $A \cap B = \emptyset$. Without loss of generality, assume that $A \neq \emptyset$ and $B \neq \emptyset$. A proposition $C \subseteq W : C \subseteq A \cup B$ if and only if:

1. $C \subseteq A$;
2. $C \subseteq B$; or
3. $C = D \cup E$ where $D \neq \emptyset, D \subseteq A$ and $E \neq \emptyset, E \subseteq B$. [1]

By definition 4.1.1:

$$\begin{aligned}
\text{bel}(A \cup B) &= \sum_{C \subseteq W : C \subseteq A \cup B} m(C) \\
&= \sum_{C \subseteq W : C \subseteq A} m(C) + \sum_{C \subseteq W : C \subseteq B} m(C) \\
&\quad + \sum_{D \subseteq W : D \neq \emptyset, D \subseteq A} \; \sum_{E \subseteq W : E \neq \emptyset, E \subseteq B} m(D \cup E) \\
&= \text{bel}(A) + \text{bel}(B) \\
&\quad + \sum_{D \subseteq W : D \neq \emptyset, D \subseteq A} \; \sum_{E \subseteq W : E \neq \emptyset, E \subseteq B} m(D \cup E)
\end{aligned}$$

m is non-negative, hence equation 58. □

Theorem 4.1.5 (Plausibility measures are sub-additive)**.**

$$\text{pl}(A \cup B) \leq \text{pl}(A) + \text{pl}(B) \tag{59}$$

Proof. Let W be a set of possible worlds and $A, B \subseteq W$ be propositions such that $A \cap B = \emptyset$. Without loss of generality, assume that $A \neq \emptyset$ and $B \neq \emptyset$. A proposition $C \subseteq W : C \cap (A \cup B) \neq \emptyset$ if and only if:

1. $C \cap A \neq \emptyset$ and $C \cap B \neq \emptyset$;

[1] I.e., if C is not a subset of A or B, then C is the union of sets D and E that are subsets of A and B respectively.

2. $C \cap A \neq \emptyset$ and $C \cap B = \emptyset$; or

3. $C \cap A = \emptyset$ and $C \cap B \neq \emptyset$.

By definition 4.1.1:

$$\begin{aligned}\mathrm{pl}(A \cup B) &= \sum_{C \subseteq W : C \cap (A \cup B) \neq \emptyset} m(C) \\ &= \sum_{C \subseteq W : C \cap A \neq \emptyset, C \cap B \neq \emptyset} m(C) \\ &\quad + \sum_{C \subseteq W : C \cap A \neq \emptyset, C \cap B = \emptyset} m(C) \\ &\quad + \sum_{C \subseteq W : C \cap A = \emptyset, C \cap B \neq \emptyset} m(C)\end{aligned}$$

Similarly:

$$\begin{aligned}\mathrm{pl}(A) + \mathrm{pl}(B) &= \sum_{C \subseteq W : C \cap A \neq \emptyset} m(C) + \sum_{C \subseteq W : C \cap B \neq \emptyset} m(C) \\ &= \sum_{C \subseteq W : C \cap A \neq \emptyset, C \cap B \neq \emptyset} m(C) + \sum_{C \subseteq W : C \cap A \neq \emptyset, C \cap B = \emptyset} m(C) \\ &\quad + \sum_{C \subseteq W : C \cap A \neq \emptyset, C \cap B \neq \emptyset} m(C) + \sum_{C \subseteq W : C \cap A = \emptyset, C \cap B \neq \emptyset} m(C) \\ &= \mathrm{pl}(A \cup B) + \sum_{C \subseteq W : C \cap A \neq \emptyset, C \cap B \neq \emptyset} m(C)\end{aligned}$$

m is non-negative, hence equation 59. □

Given one of m, bel or pl, the other two can be derived (theorem 4.1.6).

Theorem 4.1.6 (The mass function in terms of a belief measure)**.** *Let W be a set of possible worlds,* $\mathrm{bel} : 2^W \to [0,1]$ *be a belief measure, and $A \subseteq W$ be a proposition. The mass function $m : 2^W \to [0,1]$ is:*

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} \mathrm{bel}(B) \ \forall \ A \subseteq W \tag{60}$$

Proof. By induction on $|A|$. In the case that $|A| = 1$, $A = \{w\} : w \in W$ and $\mathrm{bel}(A) = m(A)$. Suppose that equation 60 holds for $|A| \leq n$. By definition 4.1.1, if $|A| = n + 1$:

$$\mathrm{bel}(A) = \sum_{B \subseteq A} m(B) = m(A) + \sum_{B \subset A} m(B)$$

If $B \subset A$, then $|B| \leq n$. By the inductive hypothesis:

$$\mathrm{bel}(A) = m(A) + \sum_{B \subset A} \sum_{C \subseteq B} (-1)^{|B-C|} \mathrm{bel}(C)$$

Therefore:

$$m(A) = \text{bel}(A) - \sum_{B \subset A} \sum_{C \subseteq B} (-1)^{|B-C|} \text{bel}(C)$$

$$= \text{bel}(A) + \sum_{B \subset A} \sum_{C \subseteq B} (-1)^{|B-C|+1} \text{bel}(C)$$

$$= \text{bel}(A) + \sum_{C \subset A} \text{bel}(C) \sum_{B : C \subseteq B \subseteq A} (-1)^{|B-C|+1} \tag{61}$$

We have that:

- $C \subseteq B \;\Rightarrow\; |B - C| + 1 = |B| - |C| + 1$; and
- $C \subseteq B \subset A \;\Rightarrow\; 0 \leq |B| - |C| \leq |A| - |C| - 1$.

A set $B : C \subseteq B \subset A$ is generated by choosing i elements from $A \cap C^c$ and taking their union with D. There are $\binom{|A|-|C|}{i}$ ways to do this. Hence:

$$\sum_{B : C \subseteq B \subset A} (-1)^{|B-C|+1} = \sum_{i=0}^{|A|-|C|-1} \binom{|A| - |C|}{i} (-1)^{i+1}$$

$$= - \sum_{i=0}^{|A|-|C|-1} \binom{|A| - |C|}{i} (-1)^{i}$$

$$= (-1)^{|A|-|C|} - \sum_{i=0}^{|A|-|C|} \binom{|A| - |C|}{i} (-1)^{i} \tag{62}$$

By the binomial theorem $\sum_{k=0}^{n} \binom{n}{k} r^k = (1 + r)^n$:

$$\sum_{i=0}^{|A|-|C|} \binom{|A| - |C|}{i} (-1)^{i} = (1 + (-1))^{|A|-|C|} = 0 \tag{63}$$

By substituting equation 63 into equation 62:

$$\sum_{B : C \subseteq B \subset A} (-1)^{|B-C|+1} = (-1)^{|A|-|C|} \tag{64}$$

By substituting equation 64 into equation 61:

$$m(A) = \text{bel}(A) + \sum_{C \subset A} \text{bel}(C)(-1)^{|A|-|C|}$$

$$= \sum_{C \subseteq A} (-1)^{|A|-|C|} \text{bel}(C)$$

□

4.2 Imprecise probabilities

An alternative but related approach is to represent beliefs by sets of probability measures (*credal sets*).

Definition 4.2.1 (Lower and upper probability measures)**.** *Let $\mathbb{P}(K) \subseteq \mathbb{P}$ be a closed convex set of probability measures $P : 2^W \to [0,1]$ and $A \subseteq W$ be a proposition. The lower ($\underline{P}$) and upper ($\overline{P}$) probability measures are:*

$$\underline{P}(A) = \min \{P(A) : P \in \mathbb{P}(K)\}, \ \overline{P}(A) = \max \{P(A) : P \in \mathbb{P}(K)\} \tag{65}$$

Theorem 4.2.2 (Upper probability measure of complement)**.**

$$\begin{aligned} \overline{P}(A^c) &= \max \{P(A^c) : P \in \mathbb{P}(K)\} \\ &= \max \{1 - P(A) : P \in \mathbb{P}(K)\} \\ &= 1 - \min \{P(A) : P \in \mathbb{P}(K)\} \\ &= 1 - \underline{P}(A) \end{aligned} \tag{66}$$

Belief and plausibility measures in Dempster-Shafer theory are special cases of lower and upper probability measures, respectively.

Theorem 4.2.3 (Relations to belief and plausibility measures)**.** *Let W be a set of possible worlds, $A \subseteq W$ be a proposition,* $\text{bel} : 2^W \to [0,1]$ *be a belief measure, and $K = \{P(A) \geq \text{bel}(A)\}$.*

$$\text{bel}(A) = \underline{P}(A), \ \text{pl}(A) = \overline{P}(A) \ \forall \ A \subseteq W \tag{67}$$

Proof. In two parts:

1. $\text{bel}(A) \leq P(A) \leq \text{pl}(A) \ \forall \ P \in \mathbb{P}(K), A \subseteq W$

By the definition of K, $P(A) \geq \text{bel}(A)$ and $\text{bel}(A^c) \leq P(A^c) \ \forall \ P \in \mathbb{P}(K)$. Hence, $1 - P(A^c) \leq 1 - \text{bel}(A^c)$ and $P(A) \leq \text{pl}(A)$.

2. $\exists \ P \in \mathbb{P}(K) : P(A) = \text{bel}(A) \ \forall \ A \subseteq W$

For every $B \subseteq W$, choose a possible world $w_B \in B$ such that for a given $A \subseteq W$, $B \not\subseteq A \Rightarrow w_B \in A^c$. Define P in terms of the mass function m of bel:

$$P(w) = \sum_{B \subseteq W : w_B = w} m(B)$$

$m(B)$ is non-zero only for w_b, so $\sum_{w \in W} P(w) = \sum_{B \subseteq W} m(B) = 1$.

$P \in \mathbb{P}(K)$ because:

$$\begin{aligned} \text{bel}(C) &= \sum_{B \subseteq W : B \subseteq C} m(B) \ \forall \ C \subseteq W = \sum_{w \in C} \sum_{B \subseteq W : B \subseteq C, w_B = w} m(B) \\ &\leq \sum_{w \in C} \sum_{B \subseteq W : w_B = w} m(B) = \sum_{w \in C} P(w) = P(C) \end{aligned}$$

Similarly:

$$\begin{aligned} \text{bel}(A) &= \sum_{B \subseteq W : B \subseteq A} m(B) = \sum_{w \in A} \sum_{B \subseteq W : B \subseteq A, w_B = w} m(B) \\ &= \sum_{w \in A} \sum_{B \subseteq W : w_B = w} m(B) = \sum_{w \in A} P(w) = P(A) \end{aligned}$$

□

A mass function assigns 'weights' to pieces of evidence (sets of possible worlds). The definition of conditional probability (2.2.1) can be generalised to mass functions.

Definition 4.2.4 (Posterior probability given a mass function)**.** *Let W be a set of possible worlds, $A, B \subseteq W$ be propositions, $P : 2^W \to [0,1]$ be a prior probability distribution, and $m : 2^W \to [0,1]$ be a mass function. The posterior probability of A given m is:*

$$P(A \mid m) = \sum_{B \subseteq W} P(A \mid B) m(B) \tag{68}$$

$m(B) > 0 \Rightarrow P(B) > 0$, otherwise $P(A \mid m)$ is undefined. The posterior probability of w given m is:

$$P(w \mid m) = P(w) \sum_{B \subseteq W : w \in B} \frac{m(B)}{P(B)} \tag{69}$$

Theorem 4.2.5 (Relations to belief and plausibility measures)**.**

$$\mathrm{bel}(A) \leq P(A \mid m) \leq \mathrm{pl}(A) \ \forall \ A \subseteq W \tag{70}$$

Proof. By definition 4.2.4 and $B \subseteq A \Rightarrow P(A \mid B) = 1$:

$$\begin{aligned} P(A \mid m) &= \sum_{B \subseteq W} P(A \mid B) m(B) \\ &= \sum_{B \subseteq W : B \subseteq A} P(A \mid B) m(B) + \sum_{B \subseteq W : B \not\subseteq A} P(A \mid B) m(B) \\ &= \sum_{B \subseteq W : B \subseteq A} m(B) + \sum_{B \subseteq W : B \not\subseteq A} P(A \mid B) m(B) \\ &\geq \sum_{B \subseteq W : B \subseteq A} m(B) = \mathrm{bel}(A) \end{aligned}$$

$$\begin{aligned} P(A \mid m) &= 1 - P(A^c \mid m) \\ &\leq 1 - \mathrm{bel}(A^c) = \mathrm{pl}(A) \end{aligned}$$

□

If the prior probability distribution is uniform, then definition 4.2.4 is:

$$P(w \mid m) = \sum_{B \subseteq W : w \in B} \frac{m(B)}{|B|}$$

I.e., the posterior probability distribution redistributes the mass values associated with non-singleton sets uniformly to the singleton sets of their elements. This is called the *pignistic distribution* of a mass function.