Chapter 5

Reasoning with Uncertainty

In the early 1960s, researchers in applied logic assumed that theorem provers were powerful and general enough to solve practical, real-life problems. In particular, the introduction of the resolution principle by Allan Robinson led to this conviction. By and by however it became apparent that the appropriateness of mathematical logic for solving practical problems was highly overrated. One of the complications with real-life situations is that the facts and experience necessary for solving the problems often are typified by a degree of uncertainty; moreover, often the available information is imprecise and insufficient for solving the problems. Yet human experts are able to form judgements and take decisions from uncertain, incomplete and contradictory information. To be useful in an environment in which only such imprecise knowledge is available, a knowledge-based system has to capture and exploit not only the highly specialized expert knowledge, but the uncertainties that go with the represented pieces of information as well. This observation has led to the introduction of models for handling uncertain information in knowledge-based systems. Research into the representation and manipulation of uncertainty has grown into a major research area called *inexact reasoning* or *plausible reasoning*.

Probability theory is one of the oldest mathematical theories concerning uncertainty, so it is no wonder that in the early 1970s this formal theory was chosen as the first point of departure for the development of models for handling uncertain information in rule-based systems. It was soon discovered that this theory could not be applied in such a context in a straightforward manner; in Section 5.2 we shall discuss some of the problems encountered in a straightforward application of probability theory. Research then centred for a short period of time around the development of modifications of probability theory that should overcome the problems encountered and that could be applied efficiently in a rule-based environment. Several models were proposed, but neither of these presented a mathematically well-founded solution to these problems. This observation explains why we use the phrase *quasi-probabilistic models* to denote all models developed in the 1970s for rule-based systems. In this chapter, two quasi-probabilistic models will be discussed in some detail:

- the *subjective Bayesian method*, which was developed for application in the knowledge-based system PROSPECTOR;
- the *certainty factor model* which was designed by Edward Shortliffe and Bruce Buchanan for the purpose of dealing with uncertain information in MYCIN.

The treatment of these models will not only comprise a discussion of their basic notions but will also include an outline of their application in a rule-based system. In preparation for this, Section 5.1 shows which components should be present in a model for handling uncertainty in such a knowledge-based system.

The incorrectness of the quasi-probabilistic models from a mathematical point of view and an analysis of the problems the researchers were confronted with, led to a world-wide discussion concerning the appropriateness of probability theory for handling uncertain information in a knowledge-based context. This discussion has on the one hand yielded other points of departure, that is, other (more or less) mathematical foundations for models for handling uncertainty, and on the other hand new, less naive applications of probability theory. In Section 5.5 we shall present an introduction to the *Dempster-Shafer theory*, a theory which has largely been inspired by probability theory and may be considered to be an extension of it. We conclude this chapter with a discussion of two so-called *network models* which have resulted from a more recent probabilistic trend in plausible reasoning in which graphical representations of problem domains are employed.

5.1 Production rules, inference and uncertainty

In Chapter 3 we have seen that in a rule-based system the specialized domain knowledge an expert has, is modelled in production rules having the following form:

if e **then** h **fi**

The left-hand side e of such a rule is a combination of atomic conditions which are interrelated by means of the operators **and** and **or**. In the sequel such a combination of conditions will be called a (*piece of*) *evidence*. The right-hand side h of a production rule in general is a conjunction of conclusions. In this chapter we assume production rules to have just one conclusion. Notice that this restriction is not an essential one from a logical point of view. Henceforth, an atomic conclusion will be called a *hypothesis*. Furthermore, we will abstract from actions and predicates, and from variables and values, or objects, attributes, and values: conditions and conclusions will be taken to be indivisible primitives. A production rule now has the following meaning: if evidence e has been observed, then the hypothesis h is confirmed as being true.

In this Section we depart from top-down inference as the method for applying production rules, and from backward chaining as described in chapter 3, more in specific. The application of production rules as it takes place in top-down inference, may be represented graphically in a so-called *inference network*. We introduce the notion of an inference network by means of an example.

EXAMPLE 5.1

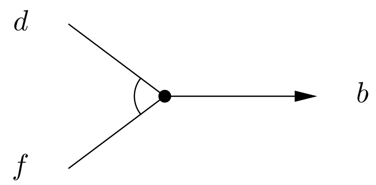
Consider the following production rules:

R_1: **if** a **and** (b **or** c) **then** h **fi**
R_2: **if** d **and** f **then** b **fi**
R_3: **if** f **or** g **then** h **fi**
R_4: **if** a **then** d **fi**

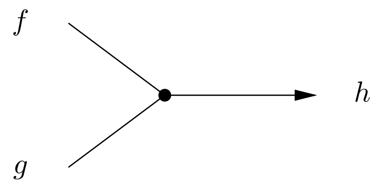
In the following, the goal for consulting a specific rule base will be called the *goal hypothesis*. We suppose that h is the goal hypothesis for consulting the set of production rules shown above. The first production rules that are selected for evaluation, are the rules R_1 and R_3. Of these, rule R_1 is evaluated first. The piece of evidence a mentioned in the left-hand side of the rule now becomes the current goal hypothesis. Since none of the production rules concludes on a, the user is requested to supply further information on a. We assume that the user confirms a being true. Subsequently, b becomes the new goal hypothesis. Since rule R_2 concludes on the hypothesis b, this rule is now selected for evaluation. The first piece of evidence mentioned in rule R_2 is d; the truth of d will be derived from rule R_4. The success of rule R_4 is depicted as follows:

$$a \longrightarrow d$$

In the evaluation of rule R_2 it remains to be examined whether or not the piece of evidence f has been observed. We assume that upon a request for further information, the user confirms the truth of f. So, rule R_2 succeeds; the success of rule R_2 is shown in the following figure:



Success of rule R_3 is depicted as follows:



The three figures shown above are the basic building blocks for constructing an inference network from a given set of production rules and a given goal hypothesis. The inference network resulting from a consultation of the four production rules of this example with h as the goal hypothesis is shown in Figure 5.1.

Up to now a production rule **if** e **then** h **fi** has been interpreted as stating: if evidence e has been observed, then the hypothesis h is confirmed as being true. In practice, however, a hypothesis seldom is confirmed to absolute certainty by the observation of a certain piece of evidence. Therefore, the notion of a production rule is extended by allowing for a *measure of uncertainty*: with the hypothesis h of the production rule **if** e **then** h **fi** a measure of uncertainty is associated indicating the degree to which h is confirmed by the observation of e.

EXAMPLE 5.2 ____

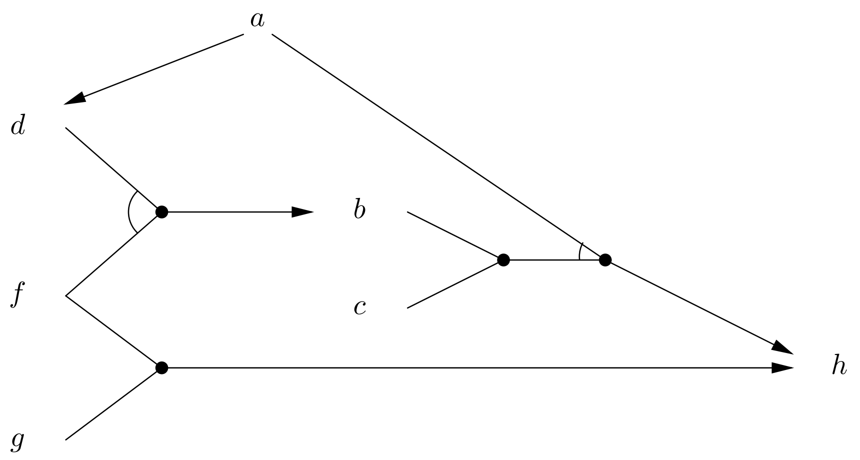
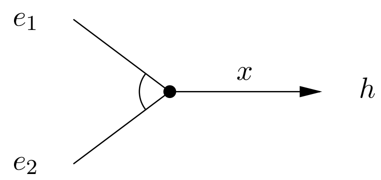


Figure 5.1: An inference network.

The measure of uncertainty x being associated with the hypothesis h in the rule **if** e_1 **and** e_2 **then** h **fi** is denoted as follows:

if e_1 **and** e_2 **then** h_x **fi**

In an inference network an associated measure of uncertainty is shown next to the arrow in the graphical representation of the rule. So, success of the production rule shown above is represented in an inference network as follows:



A model for handling uncertain information therefore provides an expert with a means for representing the uncertainties that go with the pieces of information he has specified; so, the model provides a means for *knowledge representation*.

The purpose of employing a model for dealing with uncertain information is to associate a measure of uncertainty with each conclusion the system arrives at. Such a measure of uncertainty is dependent upon the measures of uncertainty associated with the conclusions of the production rules used in deriving the final conclusion, and the measures of uncertainty the user has specified with the information he has supplied to the system. For this purpose, a model for handling uncertainty provides a means for reasoning with uncertainty, that is, it provides an *inference method*. Such an inference method consists of several components:

- Because of the way production rules of the form **if** e **then** h_y **fi** are applied during a top-down inference process, the truth of the evidence e (that is, whether or not e has actually been observed) can not always be established with absolute certainty: e may itself have been confirmed to some degree by the application of other production rules. In this case, e acts as an intermediate hypothesis that in turn is used as evidence for the confirmation of another hypothesis. The inference network shown below depicts the situation where the hypothesis e has been confirmed to the degree x on account of some prior evidence e':

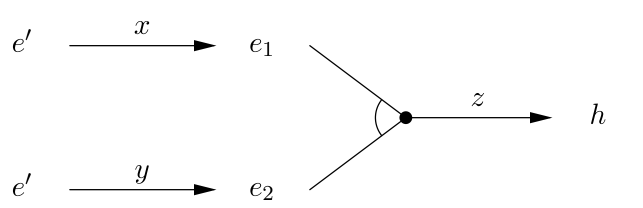
$$e' \xrightarrow{\quad x \quad} e \xrightarrow{\quad y \quad} h$$

Note that the left half of this figure shows a *compressed* inference network whereas the right half represents a single production rule. We recall that the measure of uncertainty y associated with the hypothesis h in the rule **if** e **then** h_y **fi** indicates the degree to which h is confirmed by the *actual observation*, that is, the absolute truth of e. It will be evident that in the situation shown above, we cannot simply associate the measure of uncertainty y with the hypothesis h. The actual measure of uncertainty to be associated with h depends upon y as well as on x, the measure of uncertainty associated with the evidence e used in confirming h: the uncertainty of e has to be *propagated* to h. A model for handling uncertainty provides a function for computing the actual measure of uncertainty to be associated with h on account of all prior evidence. In the sequel, such a function will be called the *combination function for (propagating) uncertain evidence*; the function will be denoted by f_{prop}. The inference network shown above can now be compressed to:

$$e' \xrightarrow{\quad f_{prop}(x, y) \quad} h$$

where e' denotes *all* prior evidence (now including e).

- The evidence e in a production rule **if** e **then** h_z **fi** in general is a combination of atomic conditions which are interrelated by means of the operators **and** and **or**. For instance, the production rule may have the form **if** e_1 **and** e_2 **then** h_z **fi** as depicted in the inference network below. Each of the constituent pieces of evidence of e may have been derived with an associated measure of uncertainty. The inference network, for example, shows that e_1 and e_2 are confirmed to the degrees x and y, respectively, on account of the prior evidence e':



To be able to apply the combination function for propagating uncertain evidence, a measure of uncertainty for e has to be computed from the measures of uncertainty that have been associated separately with the constituent pieces of evidence of e. For this purpose, a model for handling uncertainty provides two functions which will be called the *combination functions for composite hypotheses*; they will be denoted by f_{and} and f_{or}. The inference network shown above is now compressed to:

$$e' \xrightarrow{\quad f_{and}(x,y) \quad} e_1 \textbf{ and } e_2$$

- The occurrence of different production rules **if** e_i **then** h **fi** (that is, rules with different left-hand sides e_i) concluding on the same hypothesis h in the rule base, indicates that the hypothesis h may be confirmed and/or disconfirmed along different lines of reasoning. The following inference network, for example, shows the two production rules **if** e_1 **then** h_{x_2} **fi** and **if** e_2 **then** h_{y_2} **fi** concluding on the hypothesis h, the first of which uses the prior evidence e'_1 in (dis)confirming h and the second of which uses the prior evidence e'_2:

$$\begin{tikzcd}
e'_1 \arrow[r, "x_1"] & e_1 \arrow[dr, "x_2"] & \\
 & & h \\
e'_2 \arrow[r, "y_1"] & e_2 \arrow[ur, "y_2"'] &
\end{tikzcd}$$

The combination function for propagating uncertain evidence is applied to compute two *partial* measures of uncertainty x and y for h such that:

$$\begin{tikzcd}
e'_1 \arrow[dr, "x"] & \\
 & h \\
e'_2 \arrow[ur, "y"'] &
\end{tikzcd}$$

The total or *net* measure of uncertainty to be associated with h depends upon the partial measures of uncertainty that have been computed for h from the two different lines of reasoning. A model for handling uncertain information therefore provides a function for computing the net measure of uncertainty for h in the inference network shown above. Such a function will be called the *combination function for co-concluding production rules*; it will be denoted by f_{co}:

$$e' = e'_1 \textbf{ co } e'_2 \xrightarrow{\quad f_{co}(x,y) \quad} h$$

To summarize, we have introduced four combination functions:

- the function for propagating uncertain evidence: f_{prop};
- the functions for composite hypotheses: f_{and} and f_{or};
- the function for co-concluding production rules: f_{co}.

It will be evident that a model for handling uncertainty in a rule-based system has to provide fill-ins for these combination functions.

5.2 Probability theory

Probability theory is one of the earliest methods for associating with a statement a measure of uncertainty concerning its truth. In this section several notions from probability theory are introduced briefly, before we discuss the problems one encounters in applying this theory in a rule-based system in a straightforward manner.

5.2.1 The probability function

The notions that play a central role in probability theory have been developed for the description of experiments. In empirical research a more or less standard procedure is to repeatedly perform a certain experiment under essentially the same conditions. Each performance yields an *outcome* which cannot be predicted with certainty in advance. For many types of experiments, however, one is able to describe the set of all *possible* outcomes. The nonempty set of all possible outcomes of such an experiment is called its *sample space*; it is generally denoted by Ω. In the sequel, we shall only be concerned with experiments having a countable sample space.

EXAMPLE 5.3

> Consider the experiment of throwing a die. The outcome of the experiment is the number of spots up the die. The sample space of this experiment therefore consists of six elements: $\Omega = \{1, 2, 3, 4, 5, 6\}$

A subset e of the sample space Ω of a certain experiment is called an *event*. If upon performance of the experiment the outcome is in e, then it is said that the event e has occurred. In case the event e has not occurred, we use the notation $\bar{e}$, called the *complement* of e. Note that we have $\bar{e} = \Omega \setminus e$. The event that occurs if and only if both events e_1 and e_2 occur, is called the *intersection* of e_1 and e_2, and will be denoted by $e_1 \cap e_2$. The intersection of n events e_i will be denoted by

$$\bigcap_{i=1}^{n} e_i$$

The event occurring if at least one of e_1 and e_2 occurs is called the *union* of e_1 and e_2, and will be denoted by $e_1 \cup e_2$. The union of n events e_i will be denoted by

$$\bigcup_{i=1}^{n} e_i$$

EXAMPLE 5.4

> Consider the experiment of throwing a die and its associated sample space Ω once more. The subset $e_1 = \{2, 4, 6\}$ of Ω represents the event that an even number of spots has come up the die. The subset $e_2 = \bar{e}_1 = \Omega \setminus e_1 = \{1, 3, 5\}$ represents the event that an

odd number of spots has come up. The events e_1 and e_2 cannot occur simultaneously: if event e_1 occurs, that is, if an even number of spots has come up, then it is not possible that in the same throw an odd number of spots has come up. So, the event $e_1 \cap e_2$ cannot occur. Note that the event $e_1 \cup e_2$ occurs in every performance of the experiment. The subset $e_3 = \{3, 6\}$ represents the event that the number of spots that has come up is a multiple of three. Note that the events e_1 and e_3 have occurred simultaneously in case six spots are shown up the die: in that case the event $e_1 \cap e_3$ has occurred.

Definition 5.1 *The events $e_1, \ldots, e_n \subseteq \Omega$, $n \geq 1$, are called* mutually exclusive *or* disjoint events *if $e_i \cap e_j = \varnothing$, $i \neq j$, $1 \leq i, j \leq n$.*

We assume that an experiment yields an outcome independent of the outcomes of prior performances of the experiment. Now suppose that a particular experiment has been performed N times. If throughout these N performances an event e has occurred n times, the ratio $\frac{n}{N}$ is called the *relative frequency* of the occurrence of event e in N performances of the experiment. As N increases, the relative frequency of the occurrence of the event e tends to stabilize about a certain value; this value is called the *probability* that the outcome of the experiment is in e, or the probability of event e, for short.

In general, the notions of a probability and a probability function are defined axiomatically.

Definition 5.2 *Let Ω be the sample space of an experiment. If a number $P(e)$ is associated with each subset $e \subseteq \Omega$, such that*

(1) $P(e) \geq 0$,

(2) $P(\Omega) = 1$, *and*

(3) $P(\bigcup_{i=1}^{n} e_i) = \sum_{i=1}^{n} P(e_i)$, *if e_i, $i = 1, \ldots, n$, $n \geq 1$, are mutually exclusive events,*

then P is called a probability function *on the sample space Ω. For each subset $e \subseteq \Omega$, the number $P(e)$ is called the probability that event e will occur.*

Note that a probability function P on a sample space Ω is a function $P : 2^{\Omega} \to [0, 1]$.

EXAMPLE 5.5

Consider the experiment of throwing a die once more, and its associated sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The function P such that $P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = \frac{1}{6}$ is a probability function on Ω. Since the sets $\{2\}$, $\{4\}$, and $\{6\}$ are disjoint, we have according to the third axiom of the preceding definition that $P(\{2, 4, 6\}) = \frac{1}{2}$: the probability of an even number of spots coming up the die, equals $\frac{1}{2}$.

THEOREM 2 *Let Ω be the sample space of an experiment and P a probability function on Ω. Then, for each event $e \subseteq \Omega$, we have*

$$P(\bar{e}) = 1 - P(e)$$

Proof: We have $\Omega = e \cup \bar{e}$. Furthermore, $e \cap \bar{e} = \varnothing$ holds since e and $\bar{e}$ are mutually exclusive events. From the axioms 2 and 3 of the preceding definition we have that $P(\Omega) = P(e \cup \bar{e}) = P(e) + P(\bar{e}) = 1$. $\Diamond$

5.2.2 Conditional probabilities and Bayes' Theorem

We consider the case in which probability theory is applied in a medical diagnostic system. One would like to know for example the probability of the event that a specific patient has a certain disease. For many diseases, the prior probability of the disease occurring in a certain population is known. In the case of a specific patient, however, information concerning the patient's symptoms, medical history, etc. is available that might be useful in determining the probability of the presence of the disease in this specific patient.

So, in some cases we are interested only in those outcomes which are in a given nonempty subset e of the entire sample space which represents the pieces of evidence concerning the final outcome that are known in advance. Let h be the event we are interested in, that is, the hypothesis. Given that the evidence e has been observed, we now are interested in the degree to which this information influences $P(h)$, the prior probability of the hypothesis h. The probability of h given e is defined in the following definition.

Definition 5.3 *Let Ω be the sample space of a certain experiment and let P be a probability function on Ω. For each $h, e \subseteq \Omega$ with $P(e) > 0$, the* conditional probability *of h given e, denoted by $P(h \mid e)$, is defined as*

$$P(h \mid e) = \frac{P(h \cap e)}{P(e)}$$

A conditional probability $P(h \mid e)$ often is called a *posterior* probability.

The conditional probabilities given a fixed event $e \subseteq \Omega$ with $P(e) > 0$, again define a probability function on Ω since the three axioms of a probability function are satisfied:

- $P(h \mid e) = \dfrac{P(h \cap e)}{P(e)} \geq 0$, since $P(h \cap e) \geq 0$ and $P(e) > 0$;
- $P(\Omega \mid e) = \dfrac{P(\Omega \cap e)}{P(e)} = \dfrac{P(e)}{P(e)} = 1$;
- $P(\bigcup_{i=1}^{n} h_i \mid e) = \dfrac{P((\bigcup_{i=1}^{n} h_i) \cap e)}{P(e)} = \dfrac{P(\bigcup_{i=1}^{n}(h_i \cap e))}{P(e)} = \dfrac{\sum_{i=1}^{n} P(h_i \cap e)}{P(e)} =$ $\sum_{i=1}^{n} \dfrac{P(h_i \cap e)}{P(e)} = \sum_{i=1}^{n} P(h_i \mid e)$, for mutually exclusive events h_i, $i = 1, \ldots, n$, $n \geq 1$.

This probability function is called the *conditional probability function given e.*

In real-life practice, the probabilities $P(h \mid e)$ cannot always be found in the literature or obtained from statistical analysis. The conditional probabilities $P(e \mid h)$, however, often are easier to come by: in medical textbooks for example, a disease is described in terms of the signs likely to be found in a typical patient suffering from the disease. The following theorem now provides us with a method for computing the conditional probability $P(h \mid e)$ from the probabilities $P(e)$, $P(h)$, and $P(e \mid h)$; the theorem may therefore be used to reverse the 'direction' of probabilities.

THEOREM 3 (Bayes' theorem) *Let P be a probability function on a sample space Ω. For each $h, e \subseteq \Omega$ such that $P(e) > 0$ and $P(h) > 0$, we have:*

$$P(h \mid e) = \frac{P(e \mid h) \cdot P(h)}{P(e)}$$

Proof: The conditional probability of h given e is defined as

$$P(h \mid e) = \frac{P(h \cap e)}{P(e)}$$

Furthermore, we have

$$P(e \mid h) = \frac{P(e \cap h)}{P(h)}$$

So,

$$P(e \mid h) \cdot P(h) = P(h \mid e) \cdot P(e) = P(h \cap e)$$

The property stated in the theorem now follows from these observations. $\diamondsuit$

EXAMPLE 5.6

Consider the problem domain of medical diagnosis. Let h denote the hypothesis that a patient is suffering from liver cirrhosis; furthermore, let e denote the evidence that the patient has jaundice. In this case, the prior probability of liver cirrhosis, that is, $P(\textit{liver-cirrhosis})$, is known: it is the relative frequency of the disease in a particular population. If the prior probability of the occurrence of jaundice in the same population, that is, $P(\textit{jaundice})$, is likewise available and if the probability that a patient suffering from liver cirrhosis has jaundice, that is, the conditional probability $P(\textit{jaundice} \mid \textit{liver-cirrhosis})$, is known, then we can compute the probability that a patient showing signs of jaundice suffers from liver cirrhosis, that is, using Bayes' theorem we can compute the conditional probability $P(\textit{liver-cirrhosis} \mid \textit{jaundice})$. It will be evident that the last-mentioned probability is of importance in medical diagnosis.

To conclude, we define the notions of independence and conditional independence. Intuitively speaking, it seems natural to call an event h independent of an event e if $P(h \mid e) = P(h)$: the prior probability of event h is not influenced by the knowledge that event e has occurred. However, this intuitive definition of the notion of independency is not symmetrical in h and e; furthermore, the notion is defined this way only in case $P(e) > 0$. By using the definition of conditional probability and by considering the case for n events, we come to the following definition.

Definition 5.4 *The events* $e_1, \ldots, e_n \subseteq \Omega$ *are (*mutually*)* independent *if*

$$P(e_{i_1} \cap \cdots \cap e_{i_k}) = P(e_{i_1}) \cdots P(e_{i_k})$$

for each subset $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$*,* $1 \leq k \leq n$*,* $n \geq 1$*. The events* $e_1, \ldots, e_n$ *are* conditionally independent *given an event* $h \subseteq \Omega$ *if*

$$P(e_{i_1} \cap \cdots \cap e_{i_k} \mid h) = P(e_{i_1} \mid h) \cdots P(e_{i_k} \mid h)$$

for each subset $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$*.*

Note that if the events h and e are independent and if $P(e) > 0$, we have that the earlier mentioned, intuitively more appealing notion of independency

$$P(h \mid e) = \frac{P(h \cap e)}{P(e)} = \frac{P(h) \cdot P(e)}{P(e)} = P(h)$$

is satisfied.

5.2.3 Application in rule-based systems

We have mentioned before in our introduction that probability theory was chosen as the first point of departure in the pioneering work on automated reasoning under uncertainty. During the 1960s several research efforts on probabilistic reasoning were undertaken. The systems constructed in this period of time were primarily for (medical) diagnosis. Although these systems did not exhibit any intelligent reasoning behaviour, they may now be viewed as the precursors of the diagnostic systems developed in the 1970s.

Let us take a closer look at the task of diagnosis. Let $H = \{h_1, \ldots, h_n\}$ be a set of n possible hypotheses, and let $E = \{e_1, \ldots, e_m\}$ be a set of pieces of evidence which may be observed. For ease of exposition, we assume that each of the hypotheses is either true or false for a given case; equally, we assume that each of the pieces of evidence is either true (that is, it is actually observed in the given case) or false. The diagnostic task now is to find a set of hypotheses $h \subseteq H$, called the (differential) *diagnosis*, which most likely accounts for the set of observed evidence $e \subseteq E$. If we have observed a set of pieces of evidence $e \subseteq E$, then we can simply compute the conditional probabilities $P(h \mid e)$ for each subset $h \subseteq H$ and select the set $h' \subseteq H$ with the highest probability. We have mentioned before that since for real-life applications, the conditional probabilities $P(e \mid h)$ often are easier to come by than the conditional probabilities $P(h \mid e)$, generally Bayes' theorem is used for computing $P(h \mid e)$. It will be evident that the task of diagnosis in this form is computationally complex: since a diagnosis may comprise more than one hypothesis out of n possible ones, the number of diagnoses to be investigated, that is, the number of probabilities to be computed, equals 2^n. A simplifying assumption generally made in the systems for probabilistic reasoning developed in the 1960s, is that the hypotheses in H are mutually exclusive and collectively exhaustive. With this assumption, we only have to consider the n singleton hypotheses $h_i \in H$ as separate possible diagnoses. Bayes' theorem can easily be reformulated to deal with this case.

THEOREM 4 (Bayes' theorem) *Let P be a probability function on a sample space Ω. Let $h_i \subseteq \Omega$, $i = 1, \ldots, n$, $n \geq 1$, be mutually exclusive hypotheses with $P(h_i) > 0$, such that $\bigcup_{i=1}^{n} h_i = \Omega$ (that is, they are collectively exhaustive). Furthermore, let $e \subseteq \Omega$ such that $P(e) > 0$. Then, the following property holds:*

$$P(h_i \mid e) = \frac{P(e \mid h_i) \cdot P(h_i)}{\sum_{j=1}^{n} P(e \mid h_j) \cdot P(h_j)}$$

Proof: Since $h_1, \ldots, h_n$ are mutually exclusive and collectively exhaustive, we have that $P(e)$ can be written as

$$P(e) = P((\bigcup_{i=1}^{n} h_i) \cap e) = P(\bigcup_{i=1}^{n} (h_i \cap e)) = \sum_{i=1}^{n} P(h_i \cap e) = \sum_{i=1}^{n} P(e \mid h_i) \cdot P(h_i)$$

Substitution of this result in the before-mentioned form of Bayes' theorem yields the property stated in the theorem. $\Diamond$

For a successful application of Bayes' theorem in the form mentioned in the previous theorem, several conditional and prior probabilities are required. For example, conditional probabilities $P(e \mid h_i)$ for every combination of pieces of evidence $e \subseteq E$, have to be available; note that in general, these conditional probabilities $P(e \mid h_i)$ cannot be computed from their 'component' conditional probabilities $P(e_j \mid h_i)$, $e_j \in e$. It will be evident that exponentially many probabilities have to be known beforehand. Since it is hardly likely that for practical applications all these probabilities can be obtained from for example statistical analysis, a second simplifying assumption was generally made in the systems developed in the 1960s: it was assumed that the pieces of evidence $e_j \in E$ are conditionally independent given any hypothesis $h_i \in H$. Under this assumption Bayes' theorem reduces to the following form.

THEOREM 5 (Bayes' theorem) *Let P be a probability function on a sample space Ω. Let $h_i \subseteq \Omega$, $i = 1, \ldots, n$, $n \geq 1$, be mutually exclusive and collectively exhaustive hypotheses as in the previous theorem. Furthermore, let $e_{j_1}, \ldots, e_{j_k} \subseteq \Omega$, $1 \leq k \leq m$, $m \geq 1$, be pieces of evidence such that they are conditionally independent given any hypothesis h_i. Then, the following property holds:*

$$P(h_i \mid e_{j_1} \cap \cdots \cap e_{j_k}) = \frac{P(e_{j_1} \mid h_i) \cdots P(e_{j_k} \mid h_i) \cdot P(h_i)}{\sum_{i=1}^{n} P(e_{j_1} \mid h_i) \cdots P(e_{j_k} \mid h_i) \cdot P(h_i)}$$

Proof: The theorem follows immediately from the preceding theorem and the definition of conditional independence. $\diamondsuit$

It will be evident that with the two assumptions mentioned above only $m \cdot n$ conditional probabilities and $n - 1$ prior probabilities suffice for a successful use of Bayes' theorem.

The pioneering systems for probabilistic reasoning constructed in the 1960s which basically employed the last-mentioned form of Bayes' theorem, were rather small-scaled: they were devised for clear-cut problem domains with only a small number of hypotheses and restricted evidence. For these small systems, all probabilities necessary for applying Bayes' theorem were acquired from a statistical analysis of the data of several hundred sample cases. Now recall that in deriving the last-mentioned form of Bayes' theorem several assumptions were made:

- the hypotheses $h_1, \ldots, h_n$, $n \geq 1$, are mutually exclusive;
- the hypotheses $h_1, \ldots, h_n$ furthermore are collectively exhaustive, that is, $\displaystyle\bigcup_{i=1}^{n} h_i = \Omega$;
- the pieces of evidence $e_1, \ldots, e_m$, $m \geq 1$, are conditionally independent given any hypothesis h_i, $1 \leq i \leq n$.

These conditions, which have to be satisfied for a correct use of Bayes' theorem, generally are not met in practice. But, in spite of these (over-)simplifying assumptions underlying the systems from the 1960s, they performed considerably well. Nevertheless, interest in this approach to reasoning with uncertainty faded in the early 1970s. One of the reasons for this decline in interest is that the method informally sketched in the foregoing is feasible only for highly restricted problem domains: for larger domains or domains in which the above-mentioned simplifying assumptions are seriously violated, the method inevitably will become demanding, either computationally or from the point of view of obtaining the necessary probabilities: often a large number of conditional and prior probabilities is needed, thus requiring enormous amounts of experimental data.

At this stage, the first diagnostic rule-based systems began to emerge from the early artificial intelligence research efforts. As a consequence of their ability to concentrate only on those hypotheses which are suggested by the evidence, these systems in principle were capable of dealing with larger and complexer problem domains than the early probabilistic systems were. At least, they were so from a computational point of view: the problem that a large number of probabilities was required still remained. In many practical applications, the experimental data necessary for computing all probabilities required simply were not available. In devising a probabilistic reasoning component to be incorporated in a rule-based system, the artificial intelligence researchers therefore had to depart from *subjective probabilities* which had been assessed by human experts in the field. Human experts, however, often are uncertain and uncomfortable about the probabilities they are providing. The difficulty of assessing probabilities is well-known as a result of research on human decision making and judgement under uncertainty. We do not discuss this issue any further; we merely depart from the observation that domain experts generally are unable to fully and correctly specify a probability function on the problem domain. In a rule-based context, an expert now typically is asked to associate probabilities only with the production rules he has provided.

Recall that the production rule formalism is defined in terms of expressions more or less resembling logical formulas, whereas the notion of a probability function has been related to sets. Therefore, we have to have a mapping that transforms logical propositions into sets and that preserves probability, for then we have that the probability of an event is equivalent to the probability of the truth of the proposition asserting the occurrence of the event. A more or less standard translation of sets into logical formulas is the following: if Ω is a sample space, then we define for each event $e \subseteq \Omega$ a predicate e' such that $e'(x) = \mathit{true}$ if and only if $x \in e$. The intersection of two events then corresponds with the conjunction of two corresponding propositions; the union of two events translates into the disjunction of the corresponding propositions.

With each production rule **if** e **then** h **fi** an expert now associates a conditional probability $P(h \mid e)$ indicating the influence of the observation of evidence e on the prior probability $P(h)$ of the hypothesis h:

$$e \xrightarrow{P(h \mid e)} h$$

The last-mentioned form of Bayes' theorem now provides us with a method for computing the probability of a certain hypothesis when several pieces of evidence have been observed. Bayes' theorem therefore can be taken as the combination function for co-concluding production rules when probability theory is viewed as a method for handling uncertainty as discussed in Section 5.1. Consider the following inference network:

$$\begin{array}{ccc} e_1 & \xrightarrow{P(h_i \mid e_1)} & \\ \vdots & & h_i \\ e_m & \xrightarrow{P(h_i \mid e_m)} & \end{array}$$

Using Bayes' theorem we can compute the combined influence of the pieces of evidence $e_1, \ldots, e_m$ on the prior probability of the hypothesis h_i such that:

$$\bigcap_{j=1}^{m} e_j \xrightarrow{P(h_i \mid \bigcap_{j=1}^{m} e_j)} h_i$$

(Note that some prior probabilities have to be known to the system as well).

In a rule-based system, the production rules are used for pruning the search space of possible diagnoses; in this pruning process, heuristic as well as probabilistic criteria are employed. It is, therefore, necessary to compute the probabilities of all intermediate results derived using the production rules. However, these probabilities generally cannot be computed from the probabilities associated with the rules only: probability theory does not provide an explicit combination function for propagating uncertain evidence nor does it provide combination functions for composite hypotheses in terms of the available probabilities. We have suggested before that the quasi-probabilistic models do offer explicit combination functions. From the previous observation it will be evident that these functions cannot accord with the axioms of probability theory. Therefore, they can only be viewed as approximation functions rendering the models to some extent insensitive to the lack of a fully specified probability function and erroneous probability assessments.

5.3 The subjective Bayesian method

In the preceding section we have highlighted some of the problems one encounters when applying probability theory in a rule-based system. R.O. Duda, P.E. Hart, and N.J. Nilsson have recognized these problems and have developed a new method for handling uncertainty in PROSPECTOR, a knowledge-based system for assisting non-expert field geologists in exploring sites. Part of the knowledge incorporated in PROSPECTOR is represented in production rules. The model of Duda, Hart, and Nilsson is based on probability theory but provides solutions to the problems mentioned in the previous section.

5.3.1 The likelihood ratios

As has been mentioned before, the subjective Bayesian method is a modification of probability theory. However, the model uses the notion of 'odds' instead of the equivalent notion of probability.

Definition 5.5 *Let P be a probability function on a sample space Ω. Furthermore, let $h \subseteq \Omega$ such that $P(h) < 1$. The* prior odds *of the event h, denoted by $O(h)$, is defined as follows:*

$$O(h) = \frac{P(h)}{1 - P(h)}$$

Note that conversely

$$P(h) = \frac{O(h)}{1 + O(h)}$$

In probability theory the notion of conditional or posterior probability is used. The subjective Bayesian method uses the equivalent notion of posterior odds.

Definition 5.6 *Let P be a probability function on a sample space Ω. Let $h, e \subseteq \Omega$ such that $P(e) > 0$ and $P(h \mid e) < 1$. The* posterior odds *of a hypothesis h, given evidence e, denoted by $O(h \mid e)$, is defined as follows:*

$$O(h \mid e) = \frac{P(h \mid e)}{1 - P(h \mid e)}$$

We introduce another two notions: the positive and the negative likelihood ratios.

Definition 5.7 *Let P be a probability function on a sample space Ω. Furthermore, let $h, e \subseteq \Omega$ such that $0 < P(h) < 1$ and $P(e \mid \bar{h}) > 0$. The (*positive*)* likelihood ratio *λ, given h and e, is defined by*

$$\lambda = \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

The likelihood ratio λ often is called the *level of sufficiency*; it represents the degree to which the observation of evidence e influences the prior probability of hypothesis h. A likelihood ratio $\lambda > 1$ indicates that the observation of e tends to confirm the hypothesis h; a likelihood ratio $\lambda < 1$ indicates that the hypothesis $\bar{h}$ is confirmed to some degree by the observation of e, or in other words that the observation of e tends to disconfirm h. If $\lambda = 1$, then the observation of e does not influence the prior confidence in h.

Definition 5.8 *Let P be a probability function on a sample space Ω. Let $h, e \subseteq \Omega$ be such that $0 < P(h) < 1$ and $P(e \mid \bar{h}) < 1$. The (*negative*)* likelihood ratio *$\bar{\lambda}$, given h and e, is defined by*

$$\bar{\lambda} = \frac{1 - P(e \mid h)}{1 - P(e \mid \bar{h})}$$

The negative likelihood ratio $\bar{\lambda}$ often is called the *level of necessity*. A comparison of the likelihood ratios λ and $\bar{\lambda}$ shows that from $\lambda > 1$ it follows that $\bar{\lambda} < 1$, and vice versa; furthermore we have $\lambda = 1$ if and only if $\bar{\lambda} = 1$.

When applying the subjective Bayesian method in a production system, a positive likelihood ratio λ and a negative likelihood ratio $\bar{\lambda}$ have to be associated with each production rule **if** e **then** h **fi**:

$$e \xrightarrow{\lambda, \bar{\lambda}} h$$

Furthermore, the prior probabilities $P(h)$ as well as $P(e)$ have to be known to the system. Note that this information is not sufficient for uniquely defining a probability function on the sample space: the expert has provided probabilities for only a few events occurring in the specified production rules.

In the following section, in some cases the conditional probabilities $P(h \mid e)$ and $P(h \mid \bar{e})$ will be preferred to λ and $\bar{\lambda}$: we then assume that with each production rule these conditional probabilities are associated. We note that the probabilities $P(h \mid e)$ and $P(h \mid \bar{e})$ can be computed uniquely from λ, $\bar{\lambda}$, $P(h)$ and $P(e)$. The reader may for example verify that the following property holds:

$$P(e \mid h) = \lambda \cdot \frac{1 - \bar{\lambda}}{\lambda - \bar{\lambda}}$$

Bayes' theorem can subsequently be applied to compute the probability $P(h \mid e)$.

5.3.2 The combination functions

Recall that a model for dealing with uncertainty provides means for representing and reasoning with uncertainty. The purpose of applying such a model is to compute a measure of uncertainty for each goal hypothesis. If a probability function on the domain were known, then the probabilities of these goal hypotheses could simply be calculated from the probability function. However, as we have argued before, such a probability function is virtually never available in practical applications. The required probabilities therefore are approximated from the ones that actually are known to the system.

In a rule-based system using top-down inference, several intermediate hypotheses are confirmed or disconfirmed to some degree. We have seen before that these uncertain hypotheses may in turn be used as pieces of evidence in other production rules. In Section 5.1 a combination function for propagating such uncertain evidence has been introduced: the function f_{prop}. Recall that probability theory does not provide an explicit filling-in for this function f_{prop} in terms of the probabilities that are known to the system. The subjective Bayesian method, however, does provide such a combination function.

Suppose that the intermediate hypothesis e is used as evidence in confirming hypothesis h by applying the production rule **if** e **then** h **fi**. We suppose that the intermediate hypothesis e has been confirmed by the observation of some prior evidence e', and that for e the posterior probability $P(e \mid e')$ has been computed.

$$e' \xrightarrow{P(e \mid e')} e \xrightarrow{P(h \mid e),\, P(h \mid \bar{e})} h$$

After application of the rule, we are interested in the probability $P(h \mid e')$ such that

$$e' \xrightarrow{P(h \mid e')} h$$

Note that in general the probability $P(h \mid e')$ will not have been assessed by the expert and cannot be computed from the probability function P since P has not been fully specified. Therefore, it has to be approximated. In general, we have

$$\begin{aligned}
P(h \mid e') &= P(h \cap e \mid e') + P(h \cap \bar{e} \mid e') \\
&= \frac{P(h \cap e \cap e')}{P(e')} \cdot \frac{P(e \cap e')}{P(e \cap e')} + \frac{P(h \cap \bar{e} \cap e')}{P(e')} \cdot \frac{P(\bar{e} \cap e')}{P(\bar{e} \cap e')} \\
&= \frac{P(h \cap e \cap e')}{P(e \cap e')} \cdot \frac{P(e \cap e')}{P(e')} + \frac{P(h \cap \bar{e} \cap e')}{P(\bar{e} \cap e')} \cdot \frac{P(\bar{e} \cap e')}{P(e')} \\
&= P(h \mid e \cap e')P(e \mid e') + P(h \mid \bar{e} \cap e')P(\bar{e} \mid e')
\end{aligned}$$

We assume that if we know e to be absolutely true (or false), then the observations e' relevant to e do not provide any *further* information on the hypothesis h. This assumption can be taken into account into the formula given above as follows:

$$\begin{aligned}
P(h \mid e') &= P(h \mid e)P(e \mid e') + P(h \mid \bar{e})P(\bar{e} \mid e') \\
&= (P(h \mid e) - P(h \mid \bar{e})) \cdot P(e \mid e') + P(h \mid \bar{e})
\end{aligned}$$

We have that $P(h \mid e')$ is a linear interpolation function in $P(e \mid e')$ (since the function has the form $f(x) = ax + b$). In Figure 5.2 such an interpolation function for the situation of the

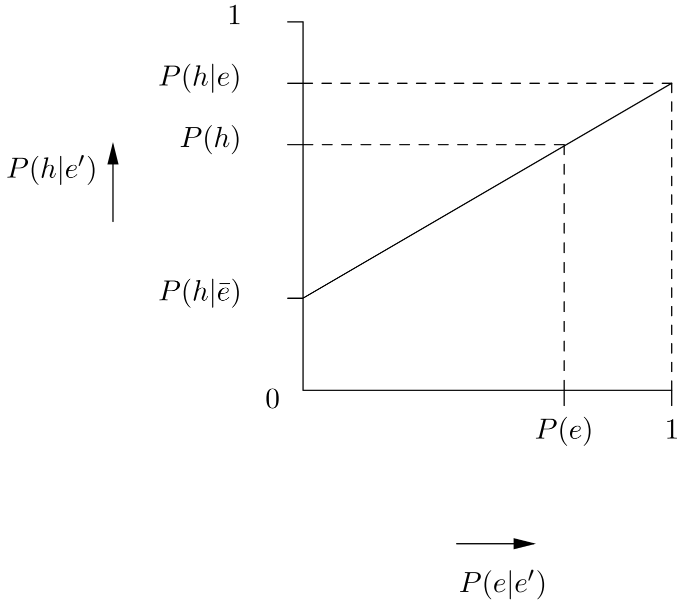


Figure 5.2: $P(h \mid e')$ as a linear interpolation function in $P(e \mid e')$.

production rule **if** e **then** h **fi** shown above, is depicted. This interpolation function has two extreme values: for $P(e \mid e') = 0$ we have the extreme value $P(h \mid e') = P(h \mid \bar{e})$, and for $P(e \mid e') = 1$ we have the extreme value $P(h \mid e') = P(h \mid e)$. For any $P(e \mid e')$ between 0 and 1 the corresponding value for $P(h \mid e')$ can be read from the figure. For instance, if evidence e' has been observed confirming e, that is, if $P(e \mid e') > P(e)$, we find that the probability of h increases from applying the production rule **if** e **then** h **fi**: $P(h \mid e') > P(h)$. Notice that this effect is exactly what is meant by the rule. In the special case where $P(e \mid e') = P(e)$, we have

$$P(h \mid e') = P(h \mid e)P(e) + P(h \mid \bar{e})P(\bar{e}) = P(h)$$

In principle, this interpolation function offers an explicit computation rule for propagating uncertain evidence. Duda, Hart, and Nilsson however have observed that when an expert is asked to assess for each rule **if** e **then** h **fi** the four probabilities $P(h), P(e), P(h \mid e)$, and $P(h \mid \bar{e})$, the specified values are likely to be *inconsistent*, in the sense that there is not an underlying actual probability function. More in specific, the relation between $P(h)$ and $P(e)$ as shown in Figure 5.2 will be violated. We show to which problems such an inconsistency may lead. Consider Figure 5.3. The assessed probabilities $P(h), P(e), P(h \mid e)$ and $P(h \mid \bar{e})$ shown in the figure are inconsistent; the consistent value for $P(e \mid e')$ corresponding with $P(h)$ is indicated as $P_c(e)$. Now suppose that evidence e' has been observed confirming e to a degree $P(e \mid e')$ such that $P(e) < P(e \mid e') < P_c(e)$. From Figure 5.3 we have that $P(h \mid e') < P(h)$. The production rule **if** e **then** h **fi** however was meant to express that confirmation of e leads to confirmation of h: due to the inconsistency the reverse has been achieved! A natural solution to this problem would be to reassess $P(e)$ by choosing $P(e) = P_c(e)$ (or, in case the assessment of $P(h)$ is less certain than the assessment of $P(e)$, to reassess $P(h)$ by choosing a consistent value for $P(h)$). The hypotheses h and e however may occur in several places in a given set of production rules and each reassessment affects all these occurrences. Reassessing prior probabilities therefore is not a feasible solution to the problem we have discussed.

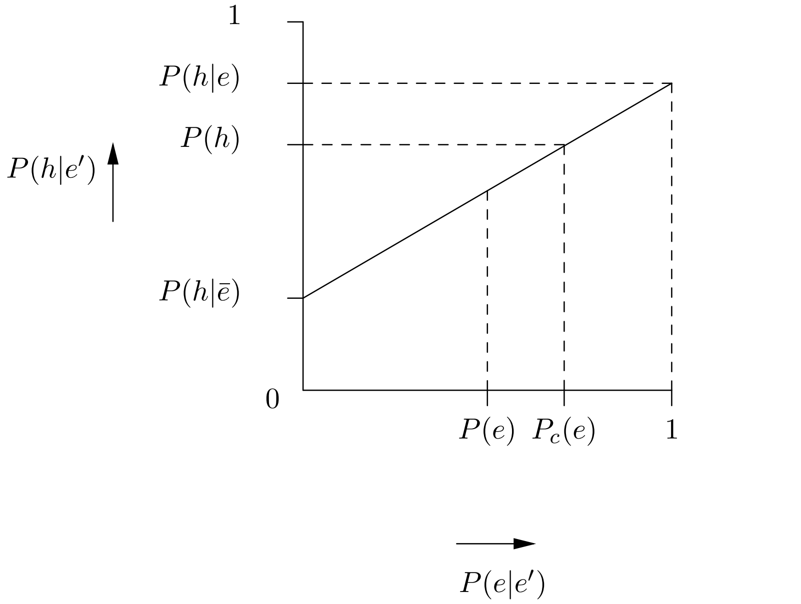


Figure 5.3: Inconsistent prior probabilities $P(h)$ and $P(e)$.

Duda, Hart, and Nilsson have developed several methods for employing inconsistently specified probabilities, one of which has been implemented as the function for propagating uncertain evidence in PROSPECTOR. The basic idea of the method that has been chosen for implementation is shown in Figure 5.4. The original interpolation function is splitted in two separate interpolation functions on the intervals $[0, P(e)]$ and $(P(e), 1]$, respectively, so as to enforce the property $P(h \mid e') = P(h)$ if $P(e \mid e') = P(e)$. Note that the closer the function value for $P(e)$ is to the value for $P(e)$ from the original interpolation function, the better the initial assessments of $P(e)$ and $P(h)$ are. The resulting interpolation function is defined as follows:

$$P(h \mid e') = \begin{cases} P(h \mid \bar{e}) + \frac{P(h)-P(h|\bar{e})}{P(e)} \cdot P(e \mid e') & \text{if } 0 \leq P(e \mid e') \leq P(e) \\ P(h) + \frac{P(h|e)-P(h)}{1-P(e)} \cdot (P(e \mid e') - P(e)) & \text{if } P(e) < P(e \mid e') \leq 1 \end{cases}$$

Recall that the conditional probabilities $P(h \mid e)$ and $P(h \mid \bar{e})$ used in this function are obtained from the likelihood ratios λ en $\bar{\lambda}$ provided by the expert.

We have mentioned before that with each production rule **if** e **then** h **fi** the two likelihood ratios λ and $\bar{\lambda}$ have been associated: λ stands for the influence of the observation of evidence e on the prior probability of the hypothesis h, and $\bar{\lambda}$ indicates the degree to which observation of $\bar{e}$ changes the probability of h. The ratios λ and $\bar{\lambda}$ can be viewed as the bounds of an interval in which lies a value indicating the degree to which evidence e, which has been (dis)confirmed to some degree by some prior evidence e', really influences the prior probability of h. This value is called the effective likelihood ratio, and will be denoted by λ'. The ratio λ' is computed from the value $P(h \mid e')$ according to the following definition.

Definition 5.9 *Let P be a probability function on a sample space Ω, and let O be the corresponding odds as defined in the foregoing. Furthermore, let $h, e' \subseteq \Omega$. The* effective likelihood ratio *λ', given h and e', is defined as follows:*

$$\lambda' = \frac{O(h \mid e')}{O(h)}$$

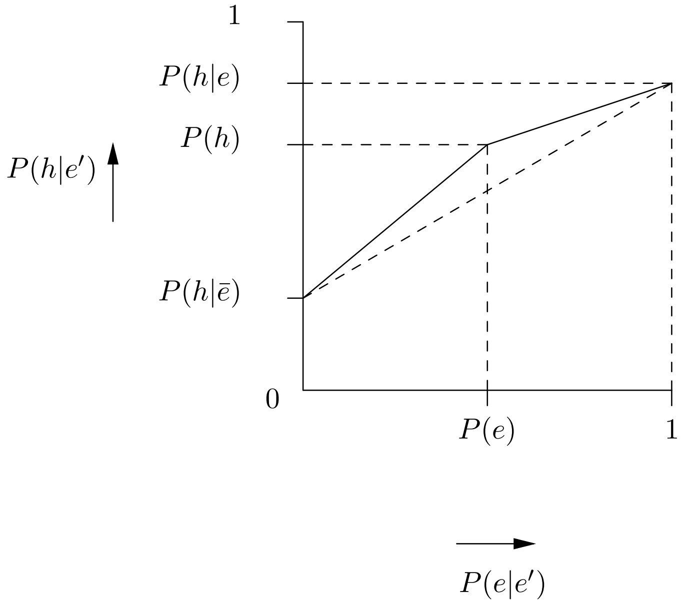
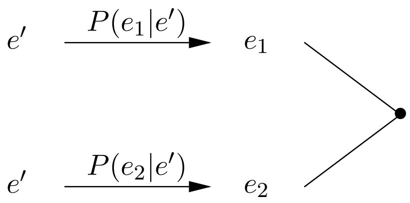


Figure 5.4: A consistent interpolation function.

The effective likelihood ratio λ' lies between λ and $\bar{\lambda}$. λ' will be closer to λ if e has been confirmed to some degree by the observation of the evidence e'; conversely, λ' will be closer to $\bar{\lambda}$ if e has been disconfirmed to some degree by the prior evidence e'.

Until now we have only considered production rules **if** e **then** h **fi** in which e is an atomic piece of evidence. In the foregoing we have seen that the condition part of a production rule may be a combination of atomic pieces of evidence which are interrelated by means of the logical operators **and** and **or**. In the inference network shown below, for example, the evidence e_1 **or** e_2 is depicted; the constituting pieces of evidence have been obtained from prior observations e':



To be able to propagate the uncertainty of the composite evidence e_1 **or** e_2, we have to know the probability $P(e_1$ **or** $e_2 \mid e')$ such that:

$$e' \xrightarrow{P(e_1 \textbf{ or } e_2|e')} e_1 \textbf{ or } e_2$$

Note that the exact probability cannot be computed from the probabilities $P(e_1 \mid e')$ and $P(e_2 \mid e')$ of the separate components. Again, we have to approximate the required probability using a combination function.

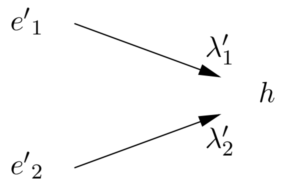
Let evidence e be composed of a number of atomic pieces of evidence e_i, $i = 1, \ldots, n, n \geq 2$, which are interrelated by means of **and** and **or**. In PROSPECTOR, the probability $P(e \mid e')$

of e given the prior observations e' is approximated from the separate probabilities $P(e_i \mid e')$ of the constituting pieces of evidence e_i in e by recursively applying the following two functions:

$$\begin{aligned} P(e_1 \textbf{ and } e_2 \mid e') &= \min\{P(e_1 \mid e'), P(e_2 \mid e')\} \\ P(e_1 \textbf{ or } e_2 \mid e') &= \max\{P(e_1 \mid e'), P(e_2 \mid e')\} \end{aligned}$$

These functions therefore fulfill the role of the combination functions for composite hypotheses, that is, of f_{and} and f_{or}, respectively. Note that the order in which the constituting pieces of evidence have been specified does not influence the resulting probability of a composite hypothesis.

The combination function which still remains to be discussed is the function for co-concluding production rules **if** e_i **then** h **fi**, that is, we still have to discuss the function f_{co}. If the pieces of evidence e_i specified in a number of co-concluding production rules have been obtained from prior observations ${e'}_i$, respectively, then the uncertainty of these pieces of evidence e_i given ${e'}_i$ can be propagated to h in the manner described above. For two co-concluding production rules, the resulting inference network is the following:



Recall that in probability theory Bayes' theorem may be used as the combination function f_{co}. In the subjective Bayesian method, Bayes' theorem is used as well, however, in a somewhat different form in terms of the odds.

THEOREM 6 *Let P be a probability function on a sample space Ω, and let O be the corresponding odds as defined above. Let $h, e \subseteq \Omega$. Furthermore, let the likelihood ratio λ be defined as above. Then, the following property holds:*

$$O(h \mid e) = \lambda \cdot O(h)$$

Proof: From Bayes' theorem we have

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)}$$

For the complement of h we have, again from Bayes' theorem,

$$P(\bar{h} \mid e) = \frac{P(e \mid \bar{h})P(\bar{h})}{P(e)}$$

Dividing the first equation by the second one results in the following equation:

$$\frac{P(h \mid e)}{P(\bar{h} \mid e)} = \frac{P(e \mid h)P(h)}{P(e \mid \bar{h})P(\bar{h})}$$

from which we have

$$\frac{P(h \mid e)}{1 - P(h \mid e)} = \frac{P(e \mid h)}{P(e \mid \bar{h})} \cdot \frac{P(h)}{1 - P(h)}$$

From this observation it follows that $O(h \mid e) = \lambda \cdot O(h)$. $\Diamond$

This alternative form of Bayes' theorem is called *odds-likelihood form* of the theorem.

The theorem stated above concerns the situation where evidence e has been obtained with absolute certainty. In case we have that e has definitely not occurred, that is, in case $\bar{e}$ has been observed with absolute certainty, we obtain a similar formula.

THEOREM 7 *Let P be a probability function on a sample space Ω, and let O be the corresponding odds as defined in the foregoing. Let $h, e \subseteq \Omega$. Furthermore, let the negative likelihood ratio $\bar{\lambda}$ be defined as above. Then, the following property holds:*

$$O(h \mid \bar{e}) = \bar{\lambda} \cdot O(h)$$

The above theorems apply to the case of a single production rule. In the situation where several production rules **if** e_i **then** h **fi** conclude on the same hypothesis h, the results from these production rules have to be combined into a single measure of uncertainty for h. Again, we first consider the case where all e_i's have been obtained with absolute certainty. It should be evident that by assuming that the e_i's are conditionally independent given h we have that the following property holds:

$$O\left(h \mid \bigcap_{i=1}^{n} e_i\right) = \prod_{i=1}^{n} \lambda_i O(h)$$

where $\lambda_i = \frac{P(e_i|h)}{P(e_i|\bar{h})}$. Similarly, for the case where all $\bar{e}_i$'s have been obtained with absolute certainty, we have:

$$O\left(h \mid \bigcap_{i=1}^{n} \bar{e}_i\right) = \prod_{i=1}^{n} \bar{\lambda}_i O(h)$$

We have argued before that in general the e_i's (or $\bar{e}_i$'s respectively) will not have been obtained with absolute certainty, but with a probability $P(e_i \mid e'_i)$ given some prior observations e'_i. From the probabilities $P(e_i \mid e'_i)$ the posterior odds $O(h \mid e'_i)$ are obtained from applying the combination function for propagating uncertain evidence. From these posterior odds we then compute the effective likelihood ratios λ'_i. Again under the assumption that the e'_i's are conditionally independent given h we obtain:

$$O\left(h \mid \bigcap_{i=1}^{n} e'_i\right) = \prod_{i=1}^{n} \lambda'_i O(h)$$

Since multiplication is commutative and associative, we have that the order in which the co-concluding production rules are applied, will be irrelevant for the resulting uncertainty for h. This finishes our discussion of the subjective Bayesian method.

5.4 The certainty factor model

The certainty factor model has been developed by E.H. Shortliffe and B.G. Buchanan for the purpose of introducing the notion of uncertainty in the MYCIN system. The development of the model was motivated, just as the subjective Bayesian method was, by the problems encountered in applying probability theory in production systems in a straightforward manner. We have suggested before that the model is unfounded from a theoretical point of view.

Nevertheless, the model has since its introduction enjoyed widespread use in rule-based systems built after MYCIN: the model has been used, and is still being used, in a large number of rule-based systems. Even though it is not well-founded, in practice it seems to behave 'satisfactorily'. The relative success of the model can be accounted for by its computational simplicity.

5.4.1 The measures of belief and disbelief

In Section 5.1 it has been argued that when modeling knowledge in production rules of the form **if** e **then** h_x **fi**, a measure of uncertainty x is associated with the hypothesis h expressing the degree to which the observation of evidence e influences the confidence in h. In developing the certainty factor model Shortliffe and Buchanan have chosen two basic measures of uncertainty: the *measure of belief* expressing the degree to which an observed piece of evidence increases the belief in a certain hypothesis, and the *measure of disbelief* expressing the degree to which an observed piece of evidence decreases the belief in a hypothesis. Although both measures are probability based, they model a notion of uncertainty conceptually different from probabilities. According to Shortliffe and Buchanan the need for new notions of uncertainty arose from their observation that an expert often was unwilling to accept the logical implications of his probabilistic statements, such as: if $P(h \mid e) = x$, then $P(\bar{h} \mid e) = 1 - x$. They state that in the mentioned case an expert would claim that 'evidence e in favour of hypothesis h should not be construed as evidence against the hypothesis as well'. The reason that the logical implication concerning $P(\bar{h} \mid e)$ may seem counterintuitive is explained by J. Pearl as follows. The phrase 'evidence e in favour of hypothesis h' is interpreted as stating an *increase* in the probability of the hypothesis from $P(h)$ to $P(h \mid e)$, with $P(h \mid e) > P(h)$: $P(h \mid e)$ is viewed relative to $P(h)$. On the other hand, in the argument of Shortliffe and Buchanan $P(\bar{h} \mid e)$ seems to be taken as an absolute probability irrespective of the prior $P(\bar{h})$. This somehow conveys the false idea that $P(\bar{h})$ increases by some positive factor. However if for example $P(\bar{h}) = 0.9$ and $P(\bar{h} \mid e) = 0.5$, then no expert will construe this considerable decrease in the probability of $\bar{h}$ as supporting the negation of h!

Anyhow, Shortliffe and Buchanan concluded from their observation that the number attached by an expert to a production rule is not a probability, but a measure of belief or disbelief in the hypothesis concerned.

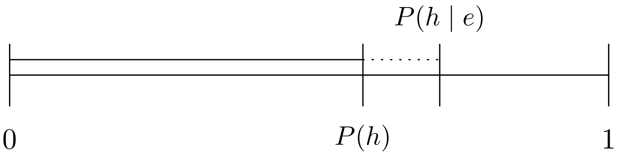
Definition 5.10 *Let P be a probability function defined on a sample space Ω, and let $h, e \subseteq \Omega$ such that $P(e) > 0$. The* measure of (increased) belief MB *is a function* $\mathrm{MB} : 2^\Omega \times 2^\Omega \to [0, 1]$*, such that*

$$\mathrm{MB}(h,e) = \begin{cases} 1 & \text{if } P(h) = 1 \\ \max\left\{0, \dfrac{P(h \mid e) - P(h)}{1 - P(h)}\right\} & \text{otherwise} \end{cases}$$

The measure of (increased) disbelief MD *is a function* $\mathrm{MD} : 2^\Omega \times 2^\Omega \to [0, 1]$*, such that*

$$\mathrm{MD}(h,e) = \begin{cases} 1 & \text{if } P(h) = 0 \\ \max\left\{0, \dfrac{P(h) - P(h \mid e)}{P(h)}\right\} & \text{otherwise} \end{cases}$$

The measure of belief can be accounted for intuitively as follows. Let us depict the prior probability of the hypothesis h, that is, $P(h)$, on a scale from 0 to 1:



The maximum amount of belief that can still be added to the prior belief in h, equals $1-P(h)$. If a piece of evidence e is observed confirming h, that is, such that $P(h \mid e) > P(h)$, then this observation results in adding the amount of belief $P(h \mid e) - P(h)$ to the prior belief in h. The belief in h therefore has been increased to the degree

$$\frac{P(h \mid e) - P(h)}{1 - P(h)}$$

The measure of disbelief can be accounted for similarly.

From the previous definition, it can readily be seen that for a given hypothesis h and a given piece of evidence e only one of the functions MB and MD attains a function value greater than zero. If $\mathrm{MB}(h,e) > 0$, we have either $P(h \mid e) - P(h) > 0$ or $P(h) = 1$. If $P(h \mid e) - P(h) > 0$ then we have $P(h) - P(h \mid e) < 0$ and consequently $\mathrm{MD}(h,e) = 0$. In case $P(h) = 1$, we have that $P(h \mid e) = 1$, hence $P(h) - P(h \mid e) = 0$ and $\mathrm{MD}(h,e) = 0$. Similarly, it can be shown that $\mathrm{MB}(h,e) = 0$ if $\mathrm{MD}(h,e) > 0$. This corresponds explicitly with the idea that a particular piece of evidence may not be used both for as well as against a hypothesis. For evidence e neither confirming nor disconfirming the hypothesis h, that is, evidence e for which $P(h \mid e) = P(h)$ holds, we have $\mathrm{MB}(h,e) = \mathrm{MD}(h,e) = 0$.

We now associate a measure of belief $\mathrm{MB}(h,e)$ and a measure of disbelief $\mathrm{MD}(h,e)$ with a hypothesis h in a production rule **if** e **then** h **fi**, as follows:

$$e \xrightarrow{\mathrm{MB}(h,e),\ \mathrm{MD}(h,e)} h$$

In this rule, the numbers $\mathrm{MB}(h,e)$ and $\mathrm{MD}(h,e)$ have the following meaning: an $\mathrm{MB}(h,e) > 0$ (and hence $\mathrm{MD}(h,e) = 0$) means that the observation of evidence e increases the confidence in h. $\mathrm{MB}(h,e) = 1$ means that the hypothesis h has been fully confirmed by e. An $\mathrm{MD}(h,e) > 0$ (and hence $\mathrm{MB}(h,e) = 0$) indicates that the observation of e tends to disconfirm the hypothesis h. Note that the measures of belief and disbelief MB and MD generally are specified by the domain expert only for a selection of the arguments in their domain. If a probability function on the domain were known, then the other function values of MB and MD could be computed using the respective definitions of these functions. However, we have argued before that such a probability function is virtually never known in practical applications. Similar to the subjective Bayesian method, the certainty factor model therefore offers a number of combination functions for approximating the function values of MB and MD that were not specified beforehand by the expert.

5.4.2 The combination functions

As we have seen before, when applying production rules various intermediate results are derived with a certain measure of uncertainty, which in turn are used as evidence in other production rules. The combination function which will be considered first, is the one for propagating such uncertainty in evidence. Suppose that an intermediate result e has been obtained from earlier evidence e' with a measure of belief $\mathrm{MB}(e, e')$ and a measure of disbelief $\mathrm{MD}(e, e')$. This e is subsequently used as evidence in the production rule **if** e **then** h **fi**:

$$e' \xrightarrow{\mathrm{MB}(e,e'),\ \mathrm{MD}(e,e')} e \xrightarrow{\mathrm{MB}(h,e), \mathrm{MD}(h,e)} h$$

Note once more that the left half of the figure shows a compressed network whereas the right half represents a single production rules. After applying the rule, we are interested in the measure of belief $\mathrm{MB}(h, e')$ and the measure of disbelief $\mathrm{MD}(h, e')$ such that:

$$e' \xrightarrow{\mathrm{MB}(h,e'), \mathrm{MD}(h,e')} h$$

The following combination functions prescribe that the measure of belief of e given e' will be used as a scaling factor for the measures of belief and disbelief associated with the production rule:

$$\begin{aligned}
\mathrm{MB}(h, e') &= \mathrm{MB}(h, e) \cdot \mathrm{MB}(e, e') \\
\mathrm{MD}(h, e') &= \mathrm{MD}(h, e) \cdot \mathrm{MB}(e, e')
\end{aligned}$$

Herein, $\mathrm{MB}(h, e)$ is the measure of belief to be assigned to the hypothesis h if the piece of evidence e has been fully confirmed; it is the measure of belief associated with h in the production rule **if** e **then** h **fi**. The meaning of $\mathrm{MD}(h, e)$ is analogous. Note that the production rule does not contribute to the belief nor to the disbelief in h if e has been disconfirmed to some extent by evidence e', in other words if the condition e has failed. The certainty factor model in this respect differs conceptually from the subjective Bayesian method.

The condition part of a production rule generally consists of a number of constituent pieces of evidence which are interrelated by means of the operators **and** and **or**. For example, the following inference network represents the composite evidence e_1 **and** e_2 where the constituent pieces of evidence e_1 and e_2 have been derived from some prior evidence e':

$$\begin{aligned}
e' &\xrightarrow{\mathrm{MB}(e_1,e'), \mathrm{MD}(e_1,e')} e_1 \\
e' &\xrightarrow{\mathrm{MB}(e_2,e'), \mathrm{MD}(e_2,e')} e_2
\end{aligned}$$

The certainty factor model comprises a number of combination functions for computing the measure of belief and the measure of disbelief for certain combinations of pieces of evidence. These combination functions are equivalent to the corresponding functions in the subjective Bayesian method:

$$\begin{aligned}
\mathrm{MB}(e_1 \textbf{ and } e_2, e') &= \min\{\mathrm{MB}(e_1, e'), \mathrm{MB}(e_2, e')\} \\
\mathrm{MB}(e_1 \textbf{ or } e_2, e') &= \max\{\mathrm{MB}(e_1, e'), \mathrm{MB}(e_2, e')\}
\end{aligned}$$

$$\begin{aligned}
\mathrm{MD}(e_1 \textbf{ and } e_2, e') &= \max\{\mathrm{MD}(e_1, e'), \mathrm{MD}(e_2, e')\} \\
\mathrm{MD}(e_1 \textbf{ or } e_2, e') &= \min\{\mathrm{MD}(e_1, e'), \mathrm{MD}(e_2, e')\}
\end{aligned}$$

The combination functions given above are commutative and associative in the first argument; so, the order in which two constituent pieces of evidence in the condition part of a production rule have been specified, has no influence on the resulting measures of belief and disbelief.

Until now, a production rule has been considered in isolation from the other production rules in a rule base. It is however possible that more than one production rule **if** e_i **then** h **fi** concludes on the same hypothesis h. Each of these different rules results in a separate measure of belief and disbelief for the same hypothesis h. We suppose that the pieces of evidence e_i specified in the co-concluding production rules have been derived from prior evidence ${e'}_i$. The uncertainty of the pieces of evidence e_i may be propagated to h in the manner described earlier in this section. For two co-concluding production rules the inference network looks as follows:

$$\begin{tikzcd}
e'_1 \arrow[dr, "\mathrm{MB}(h,e'_1),\mathrm{MD}(h,e'_1)"] & \\
& h \\
e'_2 \arrow[ur, "\mathrm{MB}(h,e'_2),\mathrm{MD}(h,e'_2)"'] &
\end{tikzcd}$$

These partial measures of belief and disbelief each contribute to the total belief and disbelief in h. The combination functions for co-concluding production rules combine these partial measures of belief and disbelief in order to obtain the total belief and disbelief in h:

$$\mathrm{MB}(h, e'_1 \ \mathbf{co}\ e'_2) \ = \begin{cases} 0 & \text{if } \mathrm{MD}(h, e'_1 \ \mathbf{co}\ e'_2) = 1 \\ \mathrm{MB}(h,e'_1) + \mathrm{MB}(h,e'_2)(1 - \mathrm{MB}(h,e'_1)) & \text{otherwise} \end{cases}$$

$$\mathrm{MD}(h, e'_1 \ \mathbf{co}\ e'_2) \ = \begin{cases} 0 & \text{if } \mathrm{MB}(h, e'_1 \ \mathbf{co}\ e'_2) = 1 \\ \mathrm{MD}(h,e'_1) + \mathrm{MD}(h,e'_2)(1 - \mathrm{MD}(h,e'_1)) & \text{otherwise} \end{cases}$$

These combination functions are commutative and associative in the second argument, so the order in which the production rules are applied has no effect on the final result.

It should be remarked that the formulas given by Shortliffe and Buchanan as shown above suggest a number of properties of the measures of belief and disbelief which do not hold in general. For instance, it is possible that the measure of belief in h as well as the measure of disbelief in h given prior evidence e' are greater than zero after applying the combination functions for co-concluding production rules, which is contradictory to the original definitions of the functions MB and MD. Only in a small number of special cases under rigorous conditions concerning the interrelationships between the pieces of evidence and the hypotheses, do the properties suggested in the formulas hold. In general, however, the combination functions are not correct with respect to the probabilistic foundation of the model.

5.4.3 The certainty factor function

In the original formulation of the certainty factor model, computation took place in terms of the measures of belief and disbelief; the uncertainties were propagated through the inference network obtained from top-down inference on a set of production rules by using the combination functions discussed above. Soon, however, the need arose to express the finally derived measures of belief and disbelief for a certain hypothesis in a single number. For this purpose, Shortliffe and Buchanan have introduced a new measure derived from the two basic ones mentioned: the certainty factor function.

Definition 5.11 *Let Ω be a sample space, and let $h, e \subseteq \Omega$. Let* MB *and* MD *be defined as in Section 5.4.1. The* certainty factor function CF *is a function* $\mathrm{CF} : 2^\Omega \times 2^\Omega \rightarrow [-1, 1]$, *such that:*

$$\mathrm{CF}(h, e) = \frac{\mathrm{MB}(h, e) - \mathrm{MD}(h, e)}{1 - \min\{\mathrm{MB}(h, e), \mathrm{MD}(h, e)\}}$$

The 'scaling factor' $1 - \min\{\mathrm{MB}(h, e), \mathrm{MD}(h, e)\}$ has been incorporated into the model for pragmatic reasons. This scaling factor has no influence on the certainty factor when considering only one piece of evidence, since then we have $1 - \min\{\mathrm{MB}(h, e), \mathrm{MD}(h, e)\} = 1$. However, when we consider more than one piece of evidence or more than one hypothesis, this is not always the case as has been mentioned before.

Note that for given h and e, a certainty factor is a number between -1 and $+1$; this is contrary to the measures of belief and disbelief, each lying in the closed interval $[0, 1]$. It can easily be seen from the definition given above that a negative certainty factor indicates that the hypothesis is disconfirmed by the evidence and that a positive certainty factor indicates that the hypothesis is confirmed by the evidence. A certainty factor equal to zero indicates that the evidence does not influence the belief in the hypothesis.

In present implementations of the certainty factor model, the measures of belief and disbelief are no longer used in the computation: only the certainty factor is applied instead of the two measures of belief and disbelief $\mathrm{MB}(h, e)$ and $\mathrm{MD}(h, e)$. With each production rule **if** e **then** h **fi** now is associated a certainty factor $\mathrm{CF}(h, e)$:

$$e \xrightarrow{\mathrm{CF}(h, e)} h$$

For manipulating these certainty factors, Shortliffe and Buchanan have defined new combination functions expressed in terms of certainty factors only. A small calculation effort suffices to prove that these combination functions can be derived from the corresponding ones for the measures of belief and disbelief.

The combination function for propagating uncertain evidence is the following:

$$\mathrm{CF}(h, e') = \mathrm{CF}(h, e) \cdot \max\{0, \mathrm{CF}(e, e')\}$$

Here, $\mathrm{CF}(h, e)$ is the certainty factor associated with the hypothesis h by the production rule **if** e **then** h **fi** if the evidence e has been observed with absolute certainty; $\mathrm{CF}(e, e')$ indicates the actual confidence in e based on some prior evidence e'.

The function for combining two certainty factors $\mathrm{CF}(e_1, e')$ and $\mathrm{CF}(e_2, e')$ of two constituting pieces of evidence e_1 and e_2 to obtain a certainty factor for the conjunction e_1 **and** e_2 of these pieces of evidence is the following:

$$\mathrm{CF}(e_1 \textbf{ and } e_2, e') = \min\{\mathrm{CF}(e_1, e'), \mathrm{CF}(e_2, e')\}$$

For the disjunction of these pieces of evidence, we have the following formula:

$$\mathrm{CF}(e_1 \textbf{ or } e_2, e') = \max\{\mathrm{CF}(e_1, e'), \mathrm{CF}(e_2, e')\}$$

Finally, the combination function for combining two certainty factors $\mathrm{CF}(h, e_1')$ and $\mathrm{CF}(h, e_2')$ which have been derived from two co-concluding production rules **if** e_i **then** h **fi**, $i = 1, 2$, is as follows:

$$\mathrm{CF}(h, e'_1 \textbf{ co } e'_2) = \begin{cases} \mathrm{CF}(h, e'_1) + \mathrm{CF}(h, e'_2)(1 - \mathrm{CF}(h, e'_1)) & \text{if } \mathrm{CF}(h, e'_i > 0, i = 1, 2 \\ \dfrac{CF(h, e'_1) + \mathrm{CF}(h, e'_2)}{1 - \min\{|\mathrm{CF}(h, e'_1)|, |\mathrm{CF}(h, e'_2)|\}} & \text{if } -1 \leq \mathrm{CF}(h, e'_1) \cdot \mathrm{CF}(h, e'_2) \leq 0 \\ \mathrm{CF}(h, e'_1) + \mathrm{CF}(h, e'_2)(1 + \mathrm{CF}(h, e'_1)) & \text{if } \mathrm{CF}(h, e'_i) < 0, i = 1, 2 \end{cases}$$

The following example demonstrates how these combination functions for certainty factors can be applied.

EXAMPLE 5.7

Consider the following five production rules:

R_1 : **if** a **and** (b **or** c) **then** $h_{0.80}$ **fi**
R_2 : **if** d **and** f **then** $b_{0.60}$ **fi**
R_3 : **if** f **or** g **then** $h_{0.40}$ **fi**
R_4 : **if** a **then** $d_{0.75}$ **fi**
R_5 : **if** i **then** $g_{0.30}$ **fi**

The expert has associated with the conclusion h of rule R_1 the certainty factor CF(h, a **and** (b **or** c)) $= 0.80$, with the conclusion b of rule R_2 the certainty factor CF(b, d **and** f) $=$ 0.60, and so on. We suppose that h is the goal hypothesis. When applying backward chaining, the user will be asked to provide further information on a, c, f and i. We assume that using his prior knowledge e', the user associates the following certainty factors with his answers:

$$\begin{aligned} CF(a, e') &= 1.00 \\ CF(c, e') &= 0.50 \\ CF(f, e') &= 0.70 \\ CF(i, e') &= -0.40 \end{aligned}$$

Using backward chaining, R_1 will be the first rule selected for application. Note that this rule will eventually yield a partial certainty factor for h. It will be evident that we cannot simply associate the certainty factor 0.80 with h after application of R_1: this number only indicates the certainty of h in case of absolute certainty of a **and** (b **or** c). Recall that for computing the actual certainty of h from this rule, we first have to compute the actual certainty of a **and** (b **or** c) and then propagate it to h using the combination function for uncertain evidence. However, the actual certainty of a **and** (b **or** c) is not known: we have to compute it from the separate certainty factors for a, b, and c using the combination functions for composite hypotheses. The actual certainty factors of a and c are known: the user has specified the certainty factors 1.00 and 0.50 for these pieces of evidence. For b, however, we still have to compute a certainty factor. We select the production rule R_2 for doing so. The combination function for uncertain evidence now prescribes that we have to multiply the certainty factor 0.60 for b mentioned in the rule by the actual certainty factor of the evidence d **and** f. Again, we have to obtain separate certainty factors for d and f. The user has associated the certainty factor 0.70 with f; by applying rule R_4 we find for d the certainty factor $1.00 \cdot 0.75 = 0.75$. Using the combination function for composite hypotheses we arrive at the following certainty factor for d **and** f (we use e'_1 to denote all evidence used in this particular reasoning chain):

$$\text{CF}(d \textbf{ and } f, e'_1) = \min\{\text{CF}(d, e'_1), \text{CF}(f, e'_1)\} = 0.70$$

Subsequently, the combination function for uncertain evidence is applied to compute the actual certainty factor for b:

$$\begin{aligned}\text{CF}(b, e'_1) &= \text{CF}(b, d \textbf{ and } f) \cdot \max\{0, \text{CF}(d \textbf{ and } f, e'_1)\} = \\ &= 0.60 \cdot 0.70 = 0.42\end{aligned}$$

Recall that we had to compute certainty factors for a, b, and c separately in order to be able to compute a certainty factor for the composite evidence a **and** (b **or** c). All the required certainty factors are now available. We apply the combination function for a disjunction of hypotheses to compute:

$$\text{CF}(b \textbf{ or } c, e'_1) = \max\{\text{CF}(b, e'_1), \text{CF}(c, e'_1)\} = 0.50$$

And, subsequently, the combination function for a conjunction of hypotheses to compute:

$$\text{CF}(a \textbf{ and } (b \textbf{ or } c), e'_1) = \min\{\text{CF}(a, e'_1), \text{CF}(b \textbf{ or } c, e'_1)\} = 0.50$$

From the production rule R_1 we therefore obtain the following (partial) certainty factor for h:

$$\begin{aligned}\text{CF}(h, e'_1) &= \text{CF}(h, a \textbf{ and } (b \textbf{ or } c)) \cdot \max\{0, \text{CF}(a \textbf{ and } (b \textbf{ or } c), e'_1)\} = \\ &= 0.80 \cdot 0.50 = 0.40\end{aligned}$$

Similarly, from the other production rule concluding on h, that is, rule R_3, the following certainty factor is obtained:

$$\begin{aligned}\text{CF}(h, e'_2) &= \text{CF}(h, f \textbf{ or } g) \cdot \max\{0, \text{CF}(f \textbf{ or } g, e'_2)\} = \\ &= 0.40 \cdot 0.70 = 0.28\end{aligned}$$

In the course of this computation a certainty factor equal to zero is associated with g due to $\text{CF}(i, e') = -0.40$. The net certainty factor for h is computed from the two partial ones by applying the combination function for co-concluding production rules:

$$\begin{aligned}\text{CF}(h, e'_1 \textbf{ co } e'_2) &= \text{CF}(h, e'_1) + \text{CF}(h, e'_2) \cdot (1 - \text{CF}(h, e'_1)) = \\ &= 0.40 + 0.28 \cdot 0.60 = 0.568\end{aligned}$$

Note that this net certainty factor is greater than each of the certainty factors for h separately.

5.5 The Dempster-Shafer theory

In the 1960s, A. Dempster laid the foundations for a new mathematical theory of uncertainty; in the seventies, this theory was extended by G. Shafer to what at present is known as the *Dempster-Shafer theory*. This theory may be viewed as a generalization of probability theory. Contrary to the subjective Bayesian method and the certainty factor model, Dempster-Shafer theory has not especially been developed for reasoning with uncertainty in knowledge-based systems. Only at the beginning of the eighties, it became apparent that the theory might be suitable for such a purpose. However, the theory cannot be applied in a knowledge-based system without modification. For application in a rule-based system, for example, several combination functions are lacking. Moreover, the theory in its original form has an exponential computational complexity. For rendering it useful in the context of knowledge-based systems, therefore, several modifications of the theory have been proposed. In Sections 5.5.1 and 5.5.2 the main principles of the theory are discussed. Section 5.5.3 briefly touches upon a possible adaptation of the theory for application in a production system.

5.5.1 The probability assignment

We have mentioned above that the Dempster-Shafer theory may be viewed as a generalization of probability theory. The development of the theory has been motivated by the observation that probability theory is not able to distinguish between *uncertainty* and *ignorance* due to incompleteness of information. Recall that in probability theory, probabilities have to be associated with individual atomic hypotheses. Only if these probabilities are known, are we able to compute other probabilities of interest. In the Dempster-Shafer theory, however, it is possible to associate measures of uncertainty with sets of hypotheses, then interpreted as disjunctions, instead of with the individual hypotheses only, and nevertheless be able to make statements concerning the uncertainty of other sets of hypotheses. Note that this way, the theory is able to distinguish between uncertainty and ignorance.

EXAMPLE 5.8

Consider a house officers' practice where a patient consults his physician for chest pain, radiating to the arms and neck; the pain does not disappear in rest. In this simplified example we assume that there are only four possible disorders to be considered as a diagnosis: the patient is either suffering from a heart attack, a pericarditis, pulmonary embolism, or an aortic dissection. Heart attack and pericarditis are disorders of the heart; pulmonary embolism and aortic dissection are disorders of the blood vessels. Now suppose that we have certain clues indicating that the patient has a disorder of the heart; the strength of our belief is expressed in the number 0.4. In the Dempster-Shafer theory this number is assigned to the set *heart-attack*, *pericarditis*, viewed as the composite hypothesis *heart-attack* **or** *pericarditis*; there is no number associated with the individual hypotheses, since more specific information indicating that one of these two hypotheses is the cause of the complaints, is not available. Note that in probability theory the number 0.4 would have to be distributed over the individual hypotheses (without more information, each of the two hypotheses would be assigned the number 0.2). In that case, the false impression of more information than actually present would be given.

The strategy followed in the Dempster-Shafer theory for dealing with uncertainty roughly amounts to the following: starting with an initial set of hypotheses, due to pieces of evidence each time a measure of uncertainty is associated with certain subsets of the original set of hypotheses, until measures of uncertainty may be associated with all possible subsets on account of the combined evidence. The initial set of all hypotheses in the problem domain is called the *frame of discernment.* In such a frame of discernment the individual hypotheses are assumed to be disjoint. The impact of a piece of evidence on the confidence or belief in certain subsets of a given frame of discernment is described by means of a function which is defined below.

Definition 5.12 *Let Θ be a frame of discernment. If with each subset $x \subseteq \Theta$ a number $m(x)$ is associated such that:*

(1) $m(x) \geq 0$,

(2) $m(\varnothing) = 0$, *and*

(3) $\sum_{x \subseteq \Theta} m(x) = 1$

then m is called a basic probability assignment *on Θ. For each subset $x \subseteq \Theta$, the number $m(x)$ is called the* basic probability number *of x.*

We define another two notions.

Definition 5.13 *Let Θ be a frame of discernment and let m be a basic probability assignment on Θ. A set $x \subseteq \Theta$ is called a* focal element *in m if $m(x) > 0$. The* core *of m, denoted by $\kappa(m)$, is the set of all focal elements in m.*

Note the similarity between a basic probability assignment and a probability function. A probability function associates with each element in Θ a number from the interval [0,1] such that the sum of these numbers equals 1; a basic probability assignment associates with each element in 2^Θ a number in the interval [0,1] such that once more the sum of the numbers equals 1.

EXAMPLE 5.9

Consider the preceding medical example once more. In this example, the frame of discernment is the set $\Theta = \{\textit{heart-attack}, \textit{pericarditis}, \textit{pulmonary-embolism}, \textit{aortic-dissection}\}$. Note that each basic probability assignment on Θ assigns basic probability numbers to $2^4 = 16$ sets (including the empty set). If for a specific patient there is no evidence pointing at a certain diagnosis in particular, then the basic probability number 1 is assigned to the entire frame of discernment:

$$m_0(x) = \begin{cases} 1 & \text{if } x = \Theta \\ 0 & \text{otherwise} \end{cases}$$

Note that each proper subset of the frame of discernment gets assigned the number 0. The core of m_0 is equal to Θ. Now suppose that some evidence has become available that points to the composite hypothesis *heart-attack* **or** *pericarditis* with some certainty. Then, the subset $\{\textit{heart-attack}, \textit{pericarditis}\}$ will be assigned a basic probability number, for example 0.4. Due to lack of further information, the remaining certainty 0.6 is assigned to the entire frame of discernment:

$$m_1(x) = \begin{cases} 0.6 & \text{if } x = \Theta \\ 0.4 & \text{if } x = \{\textit{heart-attack}, \textit{pericarditis}\} \\ 0 & \text{otherwise} \end{cases}$$

The set $\{\textit{heart-attack}, \textit{pericarditis}\}$ is an element of the core of m_1. Now suppose that we furthermore have obtained some evidence against the hypothesis that our patient is suffering from pericarditis. This information can be considered as support for the hypothesis that the patient is *not* suffering from pericarditis. This latter hypothesis is equivalent to the composite hypothesis *heart-attack* **or** *pulmonary-embolism* **or** *aortic-dissection*. In consequence of this evidence, we therefore assign a basic probability number, for example 0.7, to the set *heart-attack*, *pulmonary-embolism*, *aortic-dissection*:

$$m_2(x) = \begin{cases} 0.3 & \text{if } = \Theta \\ 0.7 & \text{if } x = \{\textit{heart-attack}, \textit{pulmonary-embolism}, \textit{aortic-dissection}\} \\ 0 & \text{otherwise} \end{cases}$$

A probability number $m(x)$ expresses the confidence or belief assigned to precisely the set x: it does not express any belief in subsets of x. It will be evident, however, that the total confidence in x is not only dependent upon the confidence in x itself, but also upon the confidence assigned to subsets of x. For a given basic probability assignment, we define a new function describing the cumulative belief in a set of hypotheses.

Definition 5.14 *Let Θ be a frame of discernment, and let m be a basic probability assignment on Θ. Then, the* belief function *(or* credibility function*) corresponding to m is the function* $\text{Bel} : 2^\Theta \to [0, 1]$ *defined by*

$$\text{Bel}(x) = \sum_{y \subseteq x} m(y)$$

for each $x \subseteq \Theta$.

Several properties of this belief function can easily be proven:

- $\text{Bel}(\Theta) = 1$ since $\sum_{y \subseteq \Theta} m(y) = 1$.
- For each $x \subseteq \Theta$ containing exactly one element, we have that $\text{Bel}(x) = m(x)$.
- For each $x \subseteq \Theta$, we have $\text{Bel}(x) + \text{Bel}(\bar{x}) \leq 1$, since

$$\text{Bel}(\Theta) = \text{Bel}(x \cup \bar{x}) = \text{Bel}(x) + \text{Bel}(\bar{x}) + \sum_{\substack{x \cap y \neq \varnothing \\ \bar{x} \cap y \neq \varnothing}} m(y) = 1$$

We furthermore have the inequality $\text{Bel}(x) + \text{Bel}(y) \neq \text{Bel}(x \cup y)$ for each $x, y \in \Theta$.

We define some special belief functions. In the preceding example, we have demonstrated how complete ignorance may be expressed. Recall that a basic probability assignment describing lack of evidence had the following form:

$$m(x) = \begin{cases} 1 & \text{if } x = \Theta \\ 0 & \text{otherwise} \end{cases}$$

The belief function corresponding to such an assignment has been given a special name.

Definition 5.15 *Let Θ be a frame of discernment and let m be a basic probability assignment such that $\kappa(m) = \{\Theta\}$. The belief function corresponding to m is called a* vacuous belief function*.*

The following definition concerns belief functions corresponding to basic probability assignments of the form

$$m(x) = \begin{cases} 1 - c_1 & \text{if } x = \Theta \\ c_1 & \text{if } x = A \\ 0 & \text{otherwise} \end{cases}$$

where $A \subseteq \Theta$, and $0 \leq c_1 \leq 1$ is a constant.

Definition 5.16 *Let Θ be a frame of discernment and let m be a basic probability assignment such that $\kappa(m) = \{A, \Theta\}$ for a certain $A \subset \Theta$. The belief function corresponding to m is called a* simple support function*.*

A belief function provides for each set x only a lower bound to the 'actual' belief in x: it is also possible that belief has been assigned to a set y such that $x \subseteq y$. Therefore, in addition to the belief function the Dempster-Shafer theory defines another function corresponding to a basic probability assignment.

Definition 5.17 *Let Θ be a frame of discernment and let m be a basic probability assignment on Θ. Then, the* plausibility function *corresponding to m is the function $\mathrm{Pl} : 2\Theta \rightarrow [0, 1]$ defined by*

$$\mathrm{Pl}(x) = \sum_{x \cap y \neq \varnothing} m(y)$$

for each $x \subseteq \Theta$.

A function value $\mathrm{Pl}(x)$ indicates the total confidence *not* assigned to $\bar{x}$. So, $\mathrm{Pl}(x)$ provides an upper bound to the 'real' confidence in x. It can easily be shown that for a given basic probability assignment m, the property

$$\mathrm{Pl}(x) = 1 - \mathrm{Bel}(\bar{x})$$

for each $x \subseteq \Theta$, holds for the the belief function Bel and the plausibility function Pl corresponding to m. The difference $\mathrm{Pl}(x) - \mathrm{Bel}(x)$ indicates the confidence in the sets y for which $x \subseteq y$ and therefore expresses the uncertainty with respect to x.

Definition 5.18 *Let Θ be a frame of discernment and let m be a basic probability assignment on Θ. Let* Bel *be the belief function corresponding to m, and let* Pl *be the plausibility function corresponding to m. For each $x \subseteq \Theta$, the closed interval* $[\mathrm{Bel}(x), \mathrm{Pl}(x)]$ *is called the* belief interval *of x.*

EXAMPLE 5.10

Let Θ be a frame of discernment, and let $x \subseteq \Theta$. Now, consider a basic probability assignment m on Θ and its corresponding functions Bel and Pl.

- If $[\mathrm{Bel}(x), \mathrm{Pl}(x)] = [0, 1]$, then no information concerning x is available.
- If $[\mathrm{Bel}(x), \mathrm{Pl}(x)] = [1, 1]$, then x has been completely confirmed by m.
- If $[\mathrm{Bel}(x), \mathrm{Pl}(x)] = [0.3, 1]$, then there is some evidence in favour of the hypothesis x.
- If $[\mathrm{Bel}(x), \mathrm{Pl}(x)] = [0.15, 0.75]$, then we have evidence in favour as well as against x.

If we have $\mathrm{Pl}(x) - \mathrm{Bel}(x) = 0$ for each $x \subseteq \Theta$, then we are back at conventional probability theory. In such a case, the belief function is called a Bayesian belief function. This notion is defined more formally in the following definition.

Definition 5.19 *Let Θ be a frame of discernment and let m be a basic probability assignment such that the core of m only consists of singleton sets. The belief function corresponding to m is called a* Bayesian belief function.

5.5.2 Dempster's rule of combination

The Dempster-Shafer theory provides a function for computing from two pieces of evidence and their associated basic probability assignment a new basic probability assignment describing the combined influence of these pieces of evidence. This function is known as *Dempster's rule of combination*. The remainder of this section is devoted to an example of the use of this function. First, however, it is defined formally in the following definition.

Definition 5.20 (Dempster's rule of combination) *Let Θ be a frame of discernment, and let m_1 and m_2 be basic probability assignments on Θ. Then, $m_1 \oplus m_2$ is a function $m_1 \oplus m_2 : 2^\Theta \to [0, 1]$ such that*

(1) $m_1 \oplus m_2(\varnothing) = 0$, and

(2) for all $x \neq \varnothing$:

$$m_1 \oplus m_2(x) = \frac{\sum_{y \cap z = x} m_1(y) \cdot m_2(z)}{\sum_{y \cap z \neq \varnothing} m_1(y) \cdot m_2(z)}$$

$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ *is the function* $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 : 2^\Theta \to [0, 1]$ *defined by*

$$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2(x) = \sum_{y \subseteq x} m_1 \oplus m_2(y)$$

The usage of Dempster's rule of combination will now be illustrated by means of an example.

EXAMPLE 5.11

Consider once more the frame of discernment $\Theta = \{heart\text{-}attack, pericarditis, pulmonary\text{-}embolism, aortic\text{-}dissection\}$. Furthermore, consider the basic probability assignment m_1 obtained from the evidence that a given patient suffers from a heart attack or a pericarditis, and the basic probability assignment m_2 obtained from the evidence that the patient does not suffer from pericarditis. These functions are shown below:

$$m_1(x) = \begin{cases} 0.6 & \text{if } x = \Theta \\ 0.4 & \text{if } x = \{heart\text{-}attack, pericarditis\} \\ 0 & \text{otherwise} \end{cases}$$

$$m_2(x) = \begin{cases} 0.3 & \text{if } x = \Theta \\ 0.7 & \text{if } x = \{heart\text{-}attack, pulmonary\text{-}embolism, aortic\text{-}dissection\} \\ 0 & \text{otherwise} \end{cases}$$

From applying Dempster's rule of combination, we obtain a new basic probability assignment $m_1 \oplus m_2$, describing the combined effect of m_1 and m_2. The basic principle of this rule is demonstrated in Figure 5.5; such a figure is called an *intersection tableau.* In front of each row of the intersection tableau is specified a subset of the frame of discernment and the basic probability number assigned to it by the basic probability assignment m_1; the figure shows only those subsets having a basic probability number not equal to zero. Above the columns of the intersection tableau again all subsets of Θ are specified, but this time with their basic probability numbers according to m_2. The crossing of a row and a column now contains the intersection of the sets associated with the row and column concerned, and specifies the product of the two basic probability numbers associated with these sets. So, at the crossing of the row corresponding to the set $\{heart\text{-}attack, pericarditis\}$ having the basic probability number 0.4, and the column corresponding to the set $\{heart\text{-}attack, pulmonary\text{-}embolism, aortic\text{-}dissection\}$ with the basic probability number 0.7, we find the set $\{heart\text{-}attack\}$ with the number 0.28.

Now observe that the set $\{heart\text{-}attack\}$ is also present at other places in the tableau since there are various possibilities for choosing two sets $x, y \subseteq \Theta$ such that $x \cap y = \{heart\text{-}attack\}$. Dempster's rule of combination now sums all basic probability numbers assigned to the set $\{heart\text{-}attack\}$. The result of this computation (possibly after normalization to 1; we shall return to this shortly) is the basic probability number assigned by $m_1 \oplus m_2$ to that specific set. The intersection tableau in Figure 5.5 shows all sets having a probability number not equal to zero. So, we have obtained the following probability assignment:

$$m_1 \oplus m_2(x) = \begin{cases} 0.18 & \text{if } x = \Theta \\ 0.28 & \text{if } x = \{heart\text{-}attack\} \\ 0.12 & \text{if } x = \{heart\text{-}attack, pericarditis\} \\ 0.42 & \text{if } x = \{heart\text{-}attack, pulmonary\text{-}embolism, aortic\text{-}dissection\} \\ 0 & \text{otherwise} \end{cases}$$

m_2 / m_1	...	{*heart-attack*, *pulmonary-embolism*, *aortic-dissection*} (0.7)	...	Θ (0.3)
...				
{*heart-attack*, *pericarditis*} (0.4)		{*heart-attack*} (0.28)		{*heart-attack*, *pericarditis*} (0.12)
...				
Θ (0.6)		{*heart-attack*, *pulmonary-embolism*, *aortic-dissection*} (0.42)		Θ (0.18)

Figure 5.5: Intersection tableau for m_1 and m_2.

m_3 / m_1	...	{*pulmonary-embolism*} (0.5)	...	Θ (0.5)
...				
{*heart-attack*, *pericarditis*} (0.4)		$\varnothing$ (0.2)		{*heart-attack*, *pericarditis*} (0.2)
...				
Θ (0.6)		{*pulmonary-embolism*} (0.3)		Θ (0.3)

Figure 5.6: An *erroneous* intersection tableau for m_1 en m_3.

However, in computing the combination of the two basic probability assignments, as demonstrated above, we may encounter a problem.

Consider m_1 once more and the basic probability assignment m_3 defined by

$$m_3(x) = \begin{cases} 0.5 & \text{if } x = \Theta \\ 0.5 & \text{if } x = \{\textit{pulmonary-embolism}\} \\ 0 & \text{otherwise} \end{cases}$$

Figure 5.6 now shows an intersection tableau which has been constructed using the same procedure as before. However, in this *erroneous* intersection tableau a basic probability assignment greater than zero has been assigned to the empty set: we have that $m_1 \oplus m_3(\varnothing) = 0.2$. So, the function $m_1 \oplus m_3$ is not a basic probability assignment, since it does not satisfy the axiom $m_1 \oplus m_3(\varnothing) = 0$. Dempster's rule of combination now simply sets $m_1 \oplus m_3(\varnothing) = 0$. As a consequence, the second axiom is violated: we now have that

$$\sum_{x \subseteq \Theta} m_1 \oplus m_3(x)$$

is less than instead of equal to 1. To remedy this problem, Dempster's rule of combination divides the remaining numbers by the scaling factor

$$\sum_{x \cap y \neq \varnothing} m_1(x) \cdot m_3(y)$$

in this example the factor 0.8. The correct intersection tableau for m_1 and m_3 is depicted in Figure 5.7.

5.5.3 Application in rule-based systems

In the preceding subsections, we have paid some attention to the principle notions of the Dempster-Shafer theory. These principles have been dealt with separate from an application in a knowledge-based system since the theory in its original form is not directly applicable as a model for plausible reasoning in this context. However, in the early eighties, research was initiated to further elaborate the model to render it suitable for application in a knowledge-based system. We have mentioned before that the basic problems preventing the use of the model in rule-based systems are its computational complexity and the lack of several combination functions. In this book, we shall not discuss the complexity problem. With respect to the second problem, various ad-hoc solutions have been proposed none of which is really satisfactory. One of these ad-hoc solutions will be briefly discussed just to illustrate the problems one encounters in providing for the missing combination functions. The simple approach sketched here has been developed by M. Ishizuka for the knowledge-based system SPERIL.

We consider a production rule **if** e_1 **then** h **fi**. The Dempster-Shafer theory does not prescribe explicitly which information should be associated with the hypothesis h of this production rule. It is rather straightforward, however, to associate a basic probability assignment with the rule. If the rule **if** e_1 **then** h **fi** is meant to express that the hypothesis h is confirmed

m_3 / m_1	$\cdots$	$\{pulmonary\text{-}embolism\}$ (0.5)	$\cdots$	Θ (0.5)
$\cdots$				
$\{heart\text{-}attack, pericarditis\}$ (0.4)		$\varnothing$ (0)		$\{heart\text{-}attack, pericarditis\}$ (0.25)
$\cdots$				
Θ (0.6)		$\{pulmonary\text{-}embolism\}$ (0.375)		Θ (0.375)

Figure 5.7: The *correct* intersection tableau for m_1 and m_3.

with certainty c_1 if the evidence e_1 has been observed with absolute certainty, then a basic probability assignment m_{e_1} such that

$$m_{e_1}(x) = \begin{cases} 1 - c_1 & \text{if } x = \Theta \\ c_1 & \text{if } x = h \\ 0 & \text{otherwise} \end{cases}$$

is associated with the hypothesis of the rule. Note that the corresponding belief function Bel_{e_1} is a simple support function. So, we have

$$e_1 \xrightarrow{m_{e_1}} h$$

Recall from Section 5.1 that plausible reasoning in a rule-based system requires the presence of a number of combination functions: a combination function for propagating uncertain evidence, a combination function for co-concluding production rules, and two combination functions for composite hypotheses. In the Dempster-Shafer theory in its original form, only the combination function for co-concluding production rules is available; we shall see that Dempster's rule of combination may be viewed as such. Consider again the production rule **if** e_1 **then** h **fi** given above and its associated functions m_{e_1} en Bel_{e_1}. Furthermore, suppose that we have a second rule **if** e_2 **then** h **fi** also concerning the hypothesis h, with the following associated basic probability assignment:

$$m_{e_2}(x) = \begin{cases} 1 - c_2 & \text{if } x = \Theta \\ c_2 & \text{if } x = h \\ 0 & \text{otherwise} \end{cases}$$

This situation is shown in the following inference network:

$$
\begin{tikzcd}
e_1 \arrow[dr, "m_{e_1}"] & \\
& h \\
e_2 \arrow[ur, "m_{e_2}"'] &
\end{tikzcd}
$$

If we assume that e_1 and e_2 have been observed with complete certainty, then the basic probability assignment that will be associated with h based on e_1 and e_2 is equal to $m_{e_1} \oplus m_{e_2}$. The other three combination functions unfortunately are lacking in the Dempster-Shafer theory.

M. Ishizuka has augmented the Dempster-Shafer theory by providing combination functions for use in his system SPERIL. We first consider the combination function for propagating uncertain evidence. Suppose that we are given a production rule **if** e **then** h **fi** with which a basic probability assignment m_e has been associated. We have seen in Section 5.1, that the evidence e is not always established with complete certainty since e itself may have been derived from applying other production rules. For example, e may have been confirmed with a measure of uncertainty $\text{Bel}_{e'}(e)$ on account of some prior evidence e':

$$e' \xrightarrow{\text{Bel}_{e'}(e)} e \xrightarrow{m_e} h$$

In this situation we are interested in $\text{Bel}_{e'}(h)$, the actual measure of uncertainty of h after application of the production rule shown above. This $\text{Bel}_{e'}(h)$ may be obtained from $m_{e'}(h)$ which is computed as follows:

$$m_{e'}(h) = m_e(h) \cdot \text{Bel}_{e'}(e)$$

Note that this provides us with a combination function for uncertain evidence. The following functions are employed in SPERIL as combination functions for composite hypotheses:

$$
\begin{aligned}
\text{Bel}_{e'}(e_1 \textbf{ and } e_2) &= \min\{\text{Bel}_{e'}(e_1), \text{Bel}_{e'}(e_2)\} \\
\text{Bel}_{e'}(e_1 \textbf{ or } e_2) &= \max\{\text{Bel}_{e'}(e_1), \text{Bel}_{e'}(e_2)\}
\end{aligned}
$$

The approach to applying Dempster-Shafer theory in a rule-based setting as sketched in this section is simple, but hardly satisfying. We have mentioned before that in the recent literature, several other approaches have been proposed, none of which is really satisfactory. We chose to discuss Ishizuka's method merely because of its simplicity and its obvious similarity to the quasi-probabilistic models treated earlier in this chapter.

5.6 Bayesian networks

In the mid-1980s a new trend in probabilistic reasoning with uncertainty in knowledge-based systems became discernable taking a graphical representation of knowledge as a point of departure. We use the phrase *network models* to denote this type of model In the preceding sections, we have concentrated primarily on models for plausible reasoning that were developed especially for knowledge-based systems using production rules for knowledge representation.

In contrast, the network models depart from another knowledge-representation formalism: the so-called *Bayesian network*. Common synonyms for the formalism are: belief network, probabilistic network, Bayesian belief network, and causal probabilistic network. Informally speaking, a Bayesian network is a graphical representation of a problem domain consisting of the statistical variables discerned in the domain and their probabilistic interrelationships. The relationships between the statistical variables are quantified by means of 'local' probabilities together defining a total probability function on the variables. This section presents a brief introduction to network models. In Section 5.6.1 we shall discuss the way knowledge is represented in a Bayesian network. The Sections 5.6.3 and 5.6.4 discuss two approaches to reasoning with such a network.

5.6.1 Knowledge representation in a Bayesian network

We have mentioned before that Bayesian networks provide a formalism for representing a problem domain. A Bayesian network comprises two parts: a *qualitative representation* of the problem domain and an associated *quantitative representation*. The qualitative part takes the form of an acyclic directed graph $G = (V(G), A(G))$ where $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, is a finite set of *vertices* and $A(G)$ is a finite set of arcs (V_i, V_j), $V_i, V_j \in V(G)$. Each vertex V_i in $V(G)$ represents a statistical variable which in general can take one of a set of values. In the sequel, however, we shall assume for simplicity's sake that the statistical variables can take only one of the truth values *true* and *false*. We take an arc $(V_i, V_j) \in A(G)$ to represent a direct 'influential' or 'causal' relationship between the variables V_i and V_j: the arc (V_i, V_j) is interpreted as stating that 'V_i directly influences V_j'. Absence of an arc between two vertices means that the corresponding variables do not influence each other directly. In general, such a directed graph has to be configured by a domain expert from human judgment; hence the phrase *belief* network. We give an example of such a qualitative representation of a problem domain.

EXAMPLE 5.12

Consider the following qualitative medical information:

> Shortness-of-breath (V_7) may be due to tuberculosis (V_2), lung cancer (V_4) or bronchitis (V_5), or more than one of them. A recent visit to Asia (V_1) increases the chances of tuberculosis, while smoking (V_3) is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray (V_8) do not discriminate between lung cancer and tuberculosis (V_6), as neither does the presence or absence of shortness-of-breath.

In this information, we may discern several statistical variables; with each variable we have associated a name V_i. The information has been represented in the acyclic directed graph G shown in Figure 5.8. Each vertex in G represents one of the statistical variables, and the arcs in G represent the causal relationships between the variables. The arc between the vertices V_3 and V_4 for example represents the information that smoking may cause lung cancer. Note that although the graph only depicts direct causal relationships, we can read indirect influences from it. For example, the graph shows that V_3 influences V_7 indirectly through V_4, V_5 and V_6: smoking may cause lung cancer and bronchitis, and these may in turn cause shortness-of-breath. However, as

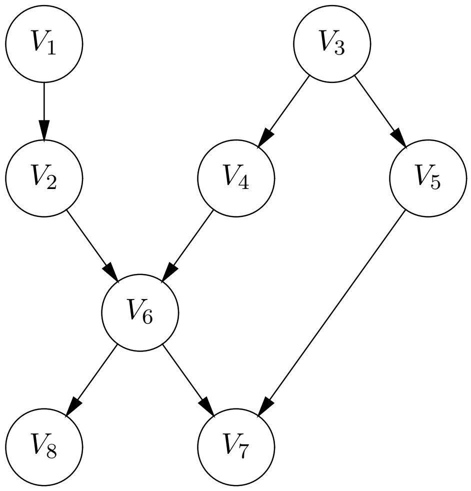


Figure 5.8: The acyclic directed graph of a Bayesian network.

soon as V_4, V_5 and V_6 are known, V_3 itself does not provide any further information concerning V_7.

The qualitative representation of the problem domain now is interpreted as the representation of all probabilistic dependency and independency relationships between the statistical variables discerned. With the graph, a domain expert associates a numerical assessment of the 'strengths' of the represented relationships in terms of a probability function P on the sample space defined by the statistical variables. Before discussing this in further detail, we introduce the notions of predecessor and successor.

Definition 5.21 *Let $G = (V(G), A(G))$ be a directed graph. Vertex $V_j \in V(G)$ is called a* successor *of vertex $V_i \in V(G)$ if there is an arc $(V_i, V_j) \in A(G)$; alternatively, vertex V_i is called a* predecessor *of vertex V_j. A vertex V_k is a neighbour of V_i if V_k is either a successor or a predecessor of V_i.*

Now, for each vertex in the graphical part of a Bayesian network, a set of (conditional) probabilities describing the influence of the values of the predecessors of the vertex on the values of the vertex itself, is specified. We shall illustrate the idea with the help of our example shortly.

We introduce some new notions and notational conventions. From now on, the variable V_i taking the truth value *true* will be denoted by v_i; the probability that the variable V_i has the value *true* will then be denoted by $P(v_i)$. We use $\neg v_i$ to denote that $V_i = \mathit{false}$; the probability that $V_i = \mathit{false}$ then is denoted by $P(\neg v_i)$. Now, let $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, again be the set of all statistical variables discerned in the problem domain. We consider a subset $V \subseteq V(G)$ with $m \geq 1$ elements. A conjunction of length m in which for each $V_i \in V$ either v_i or $\neg v_i$ occurs, is called a *configuration* of V. The conjunction $v_1 \wedge \neg v_2 \wedge v_3$ is an example of a configuration of the set $V = \{V_1, V_2, V_3\}$. The conjunction of length m in which each $V_i \in V$ is named only, that is, specified without its value, is called the *configuration template* of V. For example, the configuration template of $V = \{V_1, V_2, V_3\}$ is $V_1 \wedge V_2 \wedge V_3$. Note that we can obtain the configuration $v_1 \wedge \neg v_2 \wedge v_3$ from the template $V_1 \wedge V_2 \wedge V_3$ by filling in v_1, $\neg v_2$, and v_3 for the variables V_1, V_2, and V_3, respectively. In fact, every possible configuration of a set V can be obtained from its template by filling in proper values for the variables occurring in the template.

We return to the quantitative part of a Bayesian network. With each variable, that is, with each vertex $V_i \in V(G)$ in the qualitative part of the belief network, a domain expert associates conditional probabilities $P(v_i \mid c)$ for all configurations c of the set of predecessors of V_i in the graph. Note that for a vertex with m incoming arcs, 2^m probabilities have to be assessed; for a vertex V_i with zero predecessors, only one probability has to be specified, namely the prior probability $P(v_i)$.

EXAMPLE 5.13

Consider the medical information from the previous example and its graphical representation in Figure 5.8 once more. For example, with the vertex V_3 the domain expert associates the prior probability that a patient smokes. For the vertex V_4 two conditional probabilities have to be specified: the probability that a patient has lung cancer given the information that he smokes, that is, the probability $P(v_4 \mid v_3)$, and the probability that a non-smoker gets lung cancer, that is, the probability $P(v_4 \mid \neg v_3)$. Corresponding with the graph, a domain expert therefore has to assess the following eighteen probabilities:

$P(v_1)$
$P(v_2 \mid v_1)$ and $P(v_2 \mid \neg v_1)$
$P(v_3)$
$P(v_4 \mid v_3)$ and $P(v_4 \mid \neg v_3)$
$P(v_5 \mid v_3)$ and $P(v_5 \mid \neg v_3)$
$P(v_6 \mid v_2 \wedge v_4), P(v_6 \mid v_2 \wedge \neg v_4), P(v_6 \mid \neg v_2 \wedge v_4)$, and $P(v_6 \mid \neg v_2 \wedge \neg v_4)$
$P(v_7 \mid v_5 \wedge v_6), P(v_7 \mid v_5 \wedge \neg v_6), P(v_7 \mid \neg v_5 \wedge v_6)$, and $P(v_7 \mid \neg v_5 \wedge \neg v_6)$
$P(v_8 \mid v_6)$ and $P(v_8 \mid \neg v_6)$

Note that from these probabilities we can uniquely compute the 'complementary' probabilities; for example, we have that $P(\neg v_7 \mid v_5 \wedge v_6) = 1 - P(v_7 \mid v_5 \wedge v_6)$.

We observe that a probability function P on a sample space defined by n statistical variables $V_1, \ldots, V_n$, $n \geq 1$, is completely described by the probabilities $P(c)$ for all configurations c of $V(G) = \{V_1, \ldots, V_n\}$. The reader can easily verify that from these probabilities any other probability may be computed using the axioms mentioned in Section 5.2.1. In the sequel, therefore, we will frequently use the template $P(V_1 \wedge \cdots \wedge V_n)$ to denote a probability function: note that from this template we can obtain the probabilities $P(c)$ for all configurations c of $V(G)$, from which we can compute any probability of interest. Since there are 2^n different configurations c of $V(G)$, in theory 2^n probabilities $P(c)$ are necessary for defining a probability function. In a belief network, however, often far less probabilities suffice for doing so: an important property is that under the assumption that the graphical part of a Bayesian network represents *all* independency relationships between the statistical variables discerned, the probabilities associated with the graph provide enough information to define a unique probability function on the domain of concern. To be more precise, we have

$$P(V_1 \wedge \cdots \wedge V_n) = \prod_{i=1}^{n} P(V_i \mid C_{\rho(V_i)})$$

where $C_{\rho(V_i)}$ is the configuration template of the set $\rho(V_i)$ of predecessors of V_i. Note that the probability of any configuration of $V(G)$ can be obtained by filling in proper values for the statistical variables V_1 up to V_n inclusive and then computing the resulting product on the right-hand side from the initially assessed probabilities. We look again at our example.

EXAMPLE 5.14

Consider the previous examples once more. We have that

$$\begin{aligned} P(V_1 \wedge \cdots \wedge V_8) \;=\; & P(V_8 \mid V_6) \cdot P(V_7 \mid V_5 \wedge V_6) \cdot P(V_6 \mid V_2 \wedge V_4) \cdot \cdot P(V_5 \mid V_3) \;\cdots \\ & P(V_4 \mid V_3) \cdot P(V_3) \cdot P(V_2 \mid V_1) \cdot P(V_1) \end{aligned}$$

Note that in this example only eighteen probabilities suffice for specifying a probability function on our problem domain.

In a Bayesian network, the quantitative representation of the problem domain only comprises probabilities that involve a vertex and its predecessors in the qualitative part of the network. Note that the representation of uncertainty in such local factors closely resembles the approach followed in the quasi-probabilistic models in which uncertainty is represented in factors that are local to the production rules constituting the qualitative representation of the domain.

5.6.2 Evidence propagation in a Bayesian network

In the preceding section we have introduced the notion of a Bayesian network as a means for representing a problem domain. Such a Bayesian network may be used for reasoning with uncertainty, for example for interpreting pieces of evidence that become available during a consultation. For making probabilistic statements concerning the statistical variables discerned in the problem domain, we have to associate with a Bayesian network two methods:

- A method for computing probabilities of interest from the Bayesian network.
- A method for processing evidence, that is, a method for entering evidence into the network and subsequently computing the conditional probability function given this evidence. This process is generally called *evidence propagation*.

In the relevant literature, the emphasis lies on methods for evidence propagation; in this chapter we do so likewise.

Now recall that the probabilities associated with the graphical part of a Bayesian network uniquely define a probability function on the sample space defined by the statistical variables discerned in the problem domain. The impact of a value of a specific variable becoming known on each of the other variables, that is, the conditional probability function given the evidence, can therefore be computed from these initially assessed local probabilities. The resulting conditional probability function is often called the *updated probability function*. Calculation of a conditional probability from the initially given probabilities in a straightforward manner will generally not be restricted to computations which are local in terms of the graphical part of the Bayesian network. Furthermore, the computational complexity of such an approach is exponential in the number of variables: the method will become prohibitive for larger networks. In the literature, therefore, several less naive schemes for updating a probability function as evidence becomes available have been proposed. Although all methods build on the same notion of a Bayesian network, they differ considerably in concept and in computational complexity. All schemes proposed for evidence propagation however have two important characteristics in common:

- For propagating evidence, the graphical part of a Bayesian network is exploited more or less directly as a computational architecture.
- After a piece of evidence has been processed, again a Bayesian network results. Note that this property renders the notion of a Bayesian network invariant under evidence propagation and therefore allows for recursive application of the method for processing evidence.

In the following two sections, we shall discuss different methods for evidence propagation. In Section 5.6.3, we shall discuss the method presented by J.H. Kim and J. Pearl. In this method, computing the updated probability function after a piece of evidence has become available essentially entails each statistical variable (that is, each vertex in the graphical part of the Bayesian network) updating the probability function locally from messages it receives from its neighbours in the graph, that is, from its predecessors as well as its successors, and then in turn sending new, updated messages to them. S.L. Lauritzen and D.J. Spiegelhalter have presented another, elegant method for evidence propagation. They have observed that calculating the updated probability function after a piece of evidence has become available will generally entail going against the initially assessed 'directed' conditional probabilities. They concluded that the directed graphical representation of a Bayesian network is not suitable as an architecture for propagating evidence directly. This observation, amongst other ones, motivated an initial transformation of the Bayesian network into an undirected graphical and probabilistic representation of the problem domain. We shall see in Section 5.6.4 where this method will be discussed in some detail, that this new representation allows for an efficient method for evidence propagation in which the computations to be performed are local to small sets of variables.

5.6.3 The reasoning method of Kim and Pearl

One of the earliest methods for reasoning with a Bayesian network was proposed by J.H. Kim and J. Pearl. Their method is defined for a restricted type of Bayesian network only. It therefore is not as general as the method of Lauritzen and Spiegelhalter which will be discussed in the following section.

The method of Kim and Pearl is applicable to Bayesian networks in which the graphical part is a so-called causal polytree. A *causal polytree* is an acyclic directed graph in which between any two vertices at most one path exists. Figure 5.9 shows such a causal polytree; note that the graph shown in figure 5.8 is not a causal polytree since there exist two different paths from the vertex V_3 to the vertex V_7. For evidence propagation in their restricted type of Bayesian network, Kim and Pearl exploit the mentioned topological property of a causal polytree. Observe that from this property we have that by deleting an arbitrary arc from a causal polytree, it falls apart into two separate components. In a causal polytree G, therefore, we can identify for a vertex V_i with m neighbours, m subgraphs of G each containing a neighbour of V_i such that after removal of V_i from G there does not exist a path from one such subgraph to another one. The subgraphs corresponding with the predecessors of the

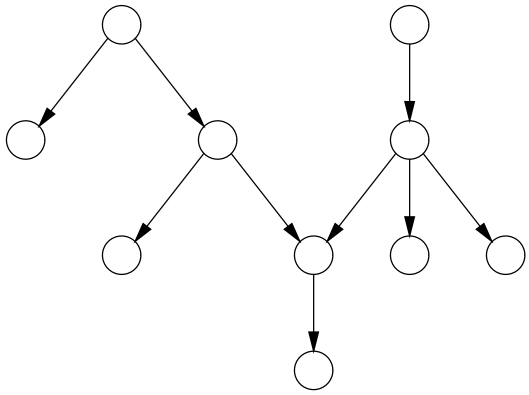


Figure 5.9: A causal polytree.

vertex will be called the *upper graphs* of V_i; the subgraphs corresponding with the successors of V_i will be called the *lower graphs* of V_i. The following example illustrates the idea. From now on, we shall restrict the discussion to this example; the reader may verify, however, that it can easily be extended to apply to more general causal polytrees.

EXAMPLE 5.15

Figure 5.10 shows a part of a causal polytree G. The vertex V_0 has the four neighbours V_1, V_2, V_3, and V_4. V_0 has two predecessors and therefore two upper graphs, which are denoted by G_1 and G_2, respectively; V_0 has also two lower graphs, denoted by G_3 and G_4. Note that there do not exist any paths between these subgraphs G_1, G_2, G_3, and G_4 other than through V_0.

So far, we have only looked at the graphical part of a Bayesian network. Recall that associated with the causal polytree we have a quantitative representation of the problem domain concerned: for each vertex, a set of local probabilities has been specified.

Let us suppose that evidence has become available that one of the statistical variables in the problem domain has adopted a specific value. This piece of evidence has to be entered into the Bayesian network in some way, and subsequently its effect on all other variables has to be computed to arrive at the updated probability function. The method for propagating evidence associated with this type of Bayesian network will be discussed shortly. First, however, we consider how probabilities of interest may be computed from the network. In doing so, we use an object-oriented style of discussion and view the causal polytree of the Bayesian network as a *computational architecture*. The vertices of the polytree are viewed as *autonomous objects* which hold some *private data* and are able to perform some computations. Recall that with each vertex is associated a set of local probabilities; these probabilities constitute the private data the object holds. The arcs of the causal polytree are taken as *communication channels*: the vertices are only able to communicate with their direct neighbours.

Now suppose that we are interested in the probabilities of the values of the variable V_0 after some evidence has been processed. It will be evident that, in terms of the graphical part of the Bayesian network, these probabilities cannot be computed from the private data the vertex holds; they are dependent upon the information from its upper and lower graphs

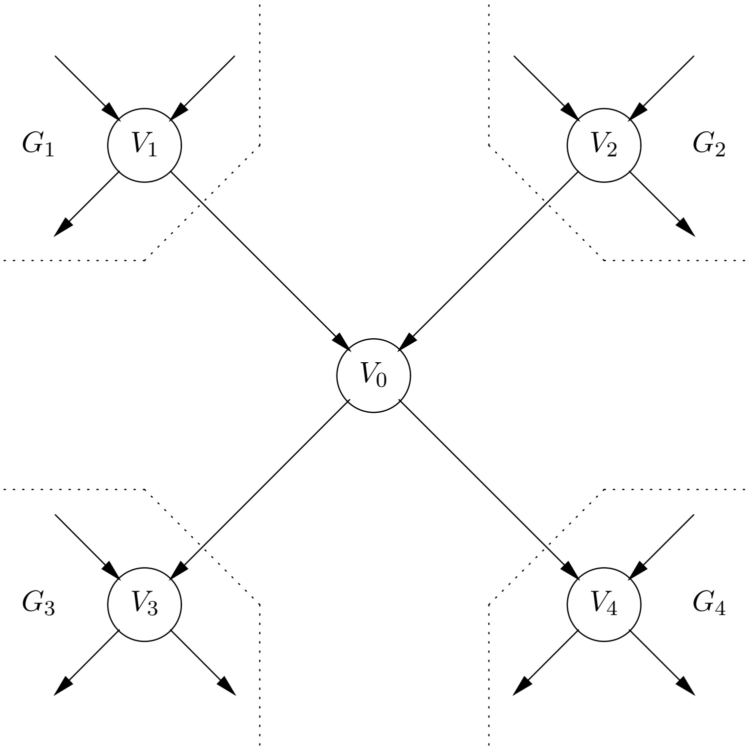


Figure 5.10: A part of a causal polytree.

as well. We shall see, however, that the neighbours of V_0 are able to provide V_0 with all information necessary for computing the probabilities of its values locally.

We introduce one more notational convention. After several pieces of evidence have been entered into the network and processed, some of the statistical variables have been *instantiated* with a value and some have not. Now, consider the configuration template $C_V(G) = V_1 \wedge \cdots \wedge V_n$ of the vertex set $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, in such a situation: we have that in the template some variables have been filled in. We shall use the notation $\tilde{c}_V(G)$ to denote the instantiated part of the template. If, for example, we have the configuration template $C = V_1 \wedge V_2 \wedge V_3$ and we know that the variable V_2 has adopted the value *true* and that the variable V_3 has the value *false*, and we do not know as yet the value of V_1, then $\tilde{c} = v_2 \wedge \neg v_3$.

We return to our example.

EXAMPLE 5.16

Consider the causal polytree from Figure 5.10 once more. We are interested in the probabilities of the values of the variable V_0. It can easily be proven, using Bayes' theorem and the independency relationships shown in the polytree, that these probabilities may be computed according to the following formula:

$$\begin{aligned} P(V_0 \mid \tilde{c}_V(G)) \quad = \quad & \alpha \cdot P(\tilde{c}_{V(G_3)} \mid V_0) \cdot P(\tilde{c}_{V(G_4)} \mid V_0) \\ & \cdot \big[P(V_0 \mid v_1 \wedge v_2) \cdot P(v_1 \mid \tilde{c}_{V(G_1)}) \cdot P(v_2 \mid \tilde{c}_{V(G_2)}) \\ & + P(V_0 \mid \neg v_1 \wedge v_2) \cdot P(\neg v_1 \mid \tilde{c}_{V(G_1)}) \cdot P(v_2 \mid \tilde{c}_{V(G_2)}) \end{aligned}$$

$$+P(V_0 \mid v_1 \wedge \neg v_2) \cdot P(v_1 \mid \tilde{c}_{V(G_1)}) \cdot P(\neg v_2 \mid \tilde{c}_{V(G_2)}) +$$
$$P(V_0 \mid \neg v_1 \wedge \neg v_2) \cdot P(\neg v_1 \mid \tilde{c}_{V(G_1)}) \cdot P(\neg v_2 \mid \tilde{c}_{V(G_2)}))]$$

where α is normalization factor chosen so as to guarantee $P(v_0 \mid \tilde{c}_V(G)) = 1 - P(\neg v_0 \mid \tilde{c}_V(G))$. We take a closer look at this formula. Note that the probabilities $P(v_0 \mid v_1 \wedge v_2)$, $P(v_0 \mid \neg v_1 \wedge v_2)$, $P(v_0 \mid v_1 \wedge \neg v_2)$, and $P(v_0 \mid \neg v_1 \wedge \neg v_2)$ necessary for computing the updated probabilities of the values of V_0 have been associated with V_0 initially: V_0 holds these probabilities as private data. So, if V_0 were to obtain the probabilities $P(\tilde{c}_{V(G_i)} \mid v_0)$ and $P(\tilde{c}_{V(G_i)} \mid \neg v_0)$ from its successors V_i, and the probabilities $Pr(v_j \mid \tilde{c}_{V(G_j)})$ and $Pr(\neg v_j \mid \tilde{c}_{V(G_j)})$ from each of its predecessors V_j, then V_0 would be able to locally compute the probabilities of its values.

In the previous example we have seen that the vertex V_0 has to receive some specific probabilities from its successors and predecessors before it is able to compute locally the probabilities of its own values. The vertex V_0 has to receive from each of its successors a so-called *diagnostic evidence parameter*: the diagnostic evidence parameter that the successor V_i sends to V_0 is a function λ_{V_i} defined by $\lambda_{V_i}(v_0) = P(\tilde{c}_{V(G_i)} \mid v_0)$ and $\lambda_{V_i}(\neg v_0) = P(\tilde{c}_{V(G_i)} \mid \neg v_0)$. The vertex V_0 furthermore has to receive from each of its predecessors a *causal evidence parameter*: the causal evidence parameter that the predecessor V_j sends to V_0 is a function π_{V_0} defined by $\pi_{V_0}(v_j) = P(v_j \mid \tilde{c}_{V(G_j)})$ and $\pi_{V_0}(\neg v_j) = P(\neg v_j \mid \tilde{c}_{V(G_j)})$. These evidence parameters may be viewed as being associated with the arcs of the causal polytree; Figure 5.11 shows the parameters associated with the causal polytree from Figure 5.10. Note that the π and λ parameters may be viewed as *messages* sent between objects.

Until now we have not addressed the question how a vertex computes the evidence parameters to be sent to its neighbours. We therefore turn our attention to evidence propagation. Suppose that evidence becomes available that a certain variable $V_i \in V(G)$ has adopted a certain value, say *true*. Informally speaking, the following happens. This evidence forces that variable V_i to update his private data: it will be evident that the updated probabilities for the values of V_i are $P(v_i) = 1$ and $P(\neg v_i) = 0$, respectively. From its local knowledge about the updated probability function, V_i then computes the proper π and λ parameters to be sent to its neighbours. V_i's neighbours subsequently are forced to update their local knowledge about the probability function and to send new parameters to their neighbours in turn. This way evidence, once entered, is spread through the Bayesian network.

EXAMPLE 5.17

Consider the causal polytree from Example 5.11 once more. The vertex V_0 computes the following causal evidence parameter to be sent to its successor V_3:

$$\begin{aligned}\pi_{V_3}(V_0) \quad = \quad & \alpha \cdot \lambda_{V_4}(V_0) \cdot [P(V_0 \mid v_1 \wedge v_2) \cdot \pi_{V_0}(v_1) \cdot \pi_{V_0}(v_2) \\ & +P(V_0 \mid \neg v_1 \wedge v_2) \cdot \pi_{V_0}(\neg v_1) \cdot \pi_{V_0}(v_2) \\ & +P(V_0 \mid v_1 \wedge \neg v_2) \cdot \pi_{V_0}(v_1) \cdot \pi_{V_0}(\neg v_2) \\ & +P(V_0 \mid \neg v_1 \wedge \neg v_2) \cdot \pi_{V_0}(\neg v_1) \cdot \pi_{V_0}(\neg v_2)]\end{aligned}$$

where α again is a normalization factor. In computing this causal evidence parameter, V_0 uses its private data and the information it obtains from its neighbours V_1, V_2, and

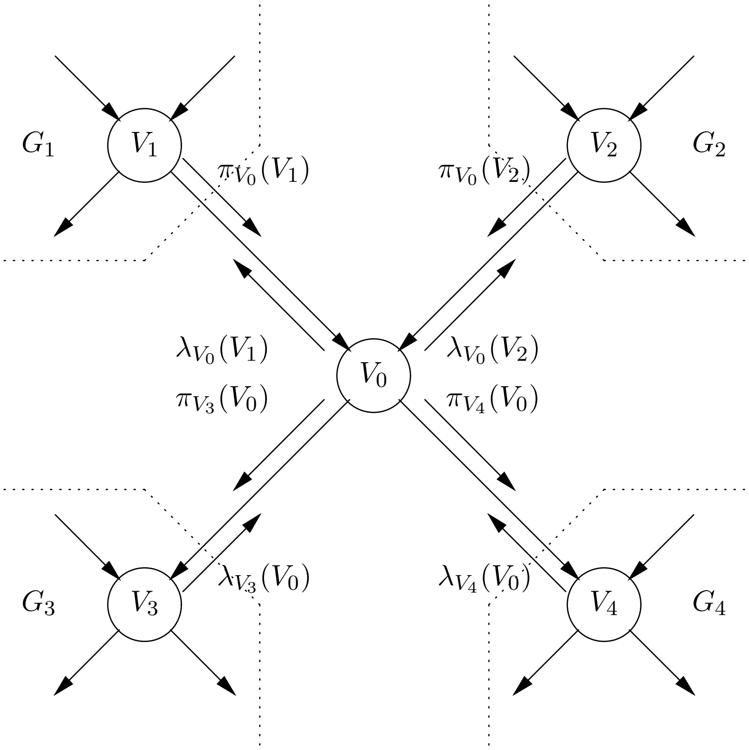


Figure 5.11: The π and λ parameters associated with the causal polytree.

V_4. Note that, if due to some new evidence for example the information $\lambda_{V_4}(V_0)$ has changed, then this change is propagated from V_4 through V_0 to V_3.

The vertex V_0 furthermore computes the following diagnostic evidence parameter to be sent to its predecessor V_1:

$$\begin{aligned}\lambda_{V_0}(V_1) \;\; = \;\; & \alpha \cdot \lambda_{V_3}(v_0) \cdot \lambda_{V_4}(v_0) \cdot [P(v_0 \mid V_1 \wedge v_2) \cdot \pi_{V_0}(v_2) \\ & + P(v_0 \mid V_1 \wedge \neg v_2) \cdot \pi_{V_0}(\neg v_2)] \\ & + \alpha \cdot \lambda_{V_3}(\neg v_0) \cdot \lambda_{V_4}(\neg v_0) \cdot [P(\neg v_0 \mid V_1 \wedge v_2) \cdot \pi_{V_0}(v_2) \\ & + P(\neg v_0 \mid V_1 \wedge \neg v_2) \cdot \pi_{V_0}(\neg v_2)]\end{aligned}$$

where α once more is a normalization factor.

We add to this example that the vertices V_i having no predecessors send a causal evidence parameter defined by $\pi_{V_j}(V_i) = P(V_i)$ to their successors V_j; furthermore, the vertices V_i having no successors initially send a diagnostic evidence parameter defined by $\lambda_{V_i}(V_j) = 1$ to their successors V_j.

We now have discussed the way a piece of evidence, once entered, is propagated through the causal polytree. We observe that any change in the joint probability distribution in response to a new piece of evidence spreads through the polytree in a single pass. This statement can readily be verified by observing that any change in the causal evidence parameter π associated with a specific arc of the causal polytree does not affect the diagnostic evidence parameter

λ on the same arc (and vice versa), since in computing the diagnostic evidence parameter $\lambda_{V_k}(V_0)$ associated with the arc (V_0, V_k) the causal evidence parameter $\pi_{V_k}(V_0)$ associated with the same arc is not used. So, in a causal polytree a perturbation is absorbed without reflection at the 'boundary' vertices, that is, vertices with either one outgoing or one incoming arc.

It remains to be discussed how a piece of evidence may be entered into the network. This is done rather elegantly: if evidence has become available that the variable V_i has the value *true* (or *false*, alternatively), then a dummy successor W of V_i is temporarily added to the polytree sending a diagnostic parameter $\lambda_W(V_i)$ to V_i such that $\lambda_W(v_i) = 1$ and $\lambda_W(\neg v)_i = 0$ (or vice versa if the value $false$ has been observed).

5.6.4 The reasoning method of Lauritzen and Spiegelhalter

In the previous section we have seen that propagating a piece of evidence concerning a specific statistical variable to the other variables in the graphical part of a Bayesian network will generally involve going against the directions of the arcs. This observation, amongst other ones, motivated S.L. Lauritzen and D.J. Spiegelhalter to transform an initially assessed Bayesian network into an equivalent undirected graphical and probabilistic representation of the problem domain. Their scheme for evidence propagation is defined on this new representation. The scheme has been inspired to a large extent by the existing statistical theory of *graphical models* (probabilistic models that can be represented by an undirected graph). In this theory, the class of so-called decomposable graphs has proven to be an important subclass of graphs. Before we define the notion of a decomposable graph, we introduce several other notions.

Definition 5.22 *Let $G = (V(G), E(G))$ be an undirected graph where $E(G)$ is a finite set of unordered pairs (V_i, V_j), $V_i, V_j \in V(G)$, called edges. A* cycle *is a path of length at least one from V_0 to V_0, $V_0 \in V(G)$. A cycle is* elementary *if all its vertices are distinct. A* chord *of an elementary cycle $V_0, V_1, \ldots, V_k = V_0$ is an edge (V_i, V_j), $i \neq (j \pm \bmod(k+1)$.*

We now are ready to define the notion of a decomposable graph.

Definition 5.23 *An undirected graph is* decomposable *if all elementary cycles of length $k >= 4$ have a chord.*

It can be shown that a probability function on such a graph may be expressed in terms of local probability functions, called *marginal* probability functions, on small sets of variables. We shall see that a representation of the problem domain in a decomposable graph and an associated representation of the probability function then allows for an efficient scheme for evidence propagation, in which the computations to be performed are local to these small sets of variables.

In order to be able to fully exploit the theory of graphical models, Lauritzen and Spiegelhalter propose a transformation of the initially assessed Bayesian network in which the graphical representation of the Bayesian network is transformed into a decomposable graph, and in which from the probabilistic part of the network a new representation of the probability function in terms of the resulting decomposable graph is obtained. The resulting representation of the problem domain is a new type of Bayesian network, which will henceforth be called a *decomposable Bayesian network*. We shall only describe the transformation of the initially assessed Bayesian network into such a decomposable Bayesian network informally.

The transformation of the original acyclic directed graph G into a decomposable graph involves three steps:

(1) Add arcs to G in such a way that no vertex in $V(G)$ has non-adjacent predecessors.

(2) Subsequently, drop the directions of the arcs.

(3) Finally, cut each elementary cycle of length four or more short by adding a chord.

It will be evident that the resulting graph is decomposable. Note that the result obtained is not unique.

EXAMPLE 5.18

Consider the Bayesian network from the example of Section 5.6.1 once more. The transformation of the graphical part of this Bayesian network into a decomposable graph is demonstrated in Figure 5.12. We consider the transformation steps in further detail. First of all, we have to add new arcs to the graph such that no vertex has non-adjacent predecessors. Now observe that in figure 5.8 the vertex V_6 has two predecessors: the vertices V_2 and V_4. Since there does not exist an arc between V_2 and V_4, we have that the predecessors of V_6 are nonadjacent. We therefore add an arc between V_2 and V_4. Note that we also have to add an arc between the vertices V_5 and V_6. Since we will drop all directions in the second transformation step, the directions of the added arcs are irrelevant. From subsequently dropping the directions of the arcs, we obtain an undirected graph. The resulting graph, however, is still not decomposable, since it has an elementary cycle of length 4 without any shortcut: V_3, V_4, V_6, V_5, V_3. We cut this cycle short by adding an edge between the vertices V_4 and V_5. Note that addition of an edge between V_3 and V_6 would have yielded a decomposable graph as well.

We now have obtained an undirected graphical representation of the problem domain. With this undirected graph, an 'undirected' representation of the probability function is associated. We confine ourselves to a discussion of this new representation, without describing how it is actually obtained from the initially assessed probabilities. It should however be evident that the new representation can be obtained from the original one, since the initial probabilities define a unique probability function.

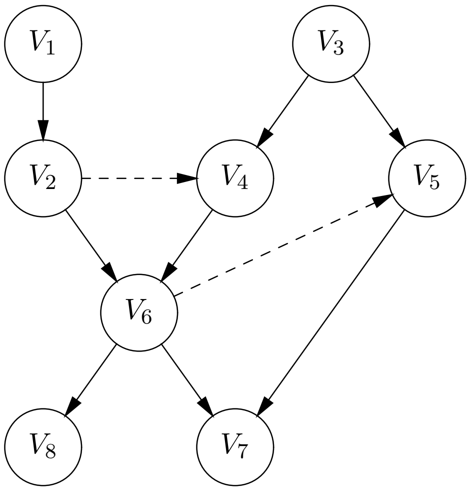
We shall see that the probability function can be expressed in terms of marginal probability functions on the cliques of the decomposable graph. We define the notion of a clique.

Definition 5.24 *Let $G = (V(G), E(G))$ be an undirected graph. A* clique *of G is a subgraph $H = (V(H), E(H))$ of G such that for any two distinct vertices $V_i, V_j \in V(H)$ we have that $(V_i, V_j) \in E(H)$. H is called a* maximal clique *of G if there does not exist a clique H' of G differing from H such that H is a subgraph of H'.*

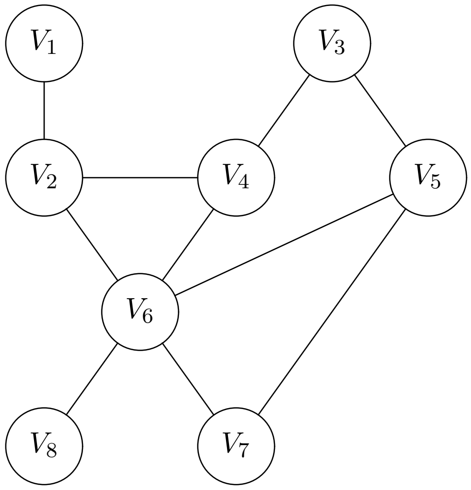
In the sequel, we shall take the word clique to mean a maximal clique.

EXAMPLE 5.19

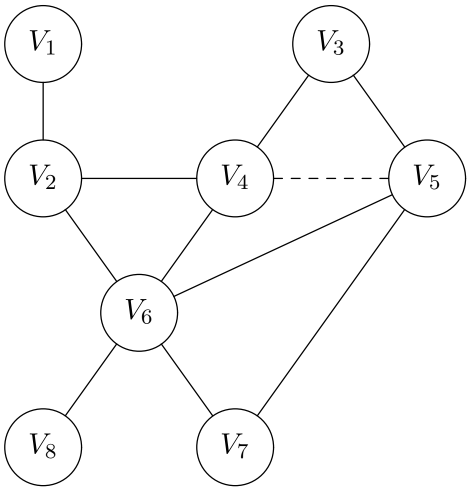
Consider the decomposable graph from Figure 5.12 once more. The reader can easily verify that this graph contains six cliques.



(a) Add arcs such that no vertex has non-adjacent predecessors



(b) Drop the directions of the arcs



(c) Cut elementary cycles short

Figure 5.12: Construction of the decomposable graph.

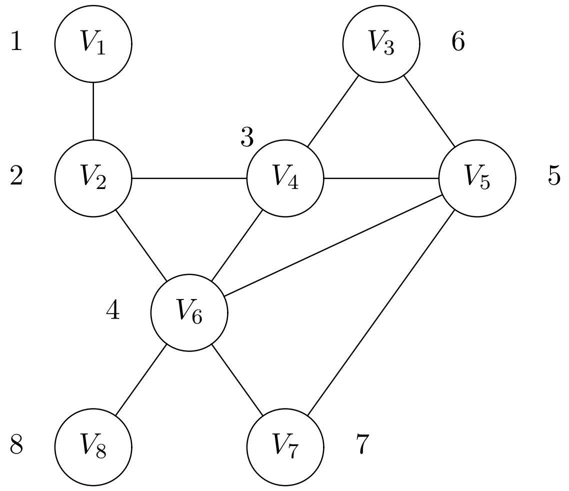


Figure 5.13: An ordering of the vertices obtained from maximum cardinality search.

To arrive at the new representation of the probability function, we obtain an ordering of the vertices and of the cliques of the decomposable graph. Its vertices are ordered as follows:

(1) Assign an arbitrary vertex the number 1.

(2) Subsequently, number the remaining vertices in increasing order such that the next number is assigned to the vertex having the largest set of previously numbered neighbours.

We say that the ordering has been obtained from *maximum cardinality search*. After the vertices of the decomposable graph have been ordered, the cliques of the graph are numbered in the order of their highest numbered vertex.

EXAMPLE 5.20

Consider the decomposable graph $G = (V(G), E(G))$ as shown in Figure 5.12 once more. The vertices of G are ordered using maximum cardinality search. An example of such an ordering is shown in Figure 5.13. The six cliques of the graph subsequently are numbered in the order of their highest numbered vertex. Let Cl_i be the clique assigned number $i, i = 1, \ldots, 6$. Then, we have obtained the following ordering (for ease of exposition we identify a clique with its vertex set):

$$\begin{aligned}
\mathrm{Cl}_1 &= \{V_1, V_2\} \\
\mathrm{Cl}_2 &= \{V_2, V_4, V_6\} \\
\mathrm{Cl}_3 &= \{V_4, V_5, V_6\} \\
\mathrm{Cl}_4 &= \{V_3, V_4, V_5\} \\
\mathrm{Cl}_5 &= \{V_5, V_6, V_7\} \\
\mathrm{Cl}_6 &= \{V_6, V_8\}
\end{aligned}$$

We consider the ordering $\mathrm{Cl}_1, \ldots, \mathrm{Cl}_m$, $m \geq 1$, of the cliques of a decomposable graph G in further detail. Let $V(\mathrm{Cl}_i)$ denote the vertex set of clique Cl_i, $i = 1, ..., m$. The ordering now has the following important property: for all $i \geq 2$ there is a $j < i$ such that $V(\mathrm{Cl}_i) \cap (V(\mathrm{Cl}_1) \cup \cdots \cup V(\mathrm{Cl}_{i-1})) \subset V(\mathrm{Cl}_j)$. In other words, the vertices a clique has in common with the lower numbered cliques are all contained in one such clique. This property is known as the *running intersection property*. This property now enables us to write the probability function on the decomposable graph as the product of the marginal probability functions on its cliques, divided by a product of the marginal probability functions on the clique intersections:

$$P(C_{V(G)}) = \prod_{i=1}^{m} \frac{P(C_{V(\mathrm{Cl}_i)})}{P(C_{S_i})}$$

where S_i is the set of vertices Cl_i has in common with the lower numbered cliques.

EXAMPLE 5.21

Consider the decomposable graph G shown in Figure 5.13 once more. The probability function on G may be expressed as

$$\begin{aligned} P(V_1 \wedge \cdots \wedge V_8) &= P(V_1 \wedge V_2) \cdot \frac{P(V_2 \wedge V_4 \wedge V_6)}{P(V_2)} \cdot \frac{P(V_4 \wedge V_5 \wedge V_6)}{P(V_4 \wedge V_6)} \\ &\quad \cdot \frac{P(V_3 \wedge V_4 \wedge V_5)}{P(V_4 \wedge V_5)} \cdot \frac{P(V_5 \wedge V_6 \wedge V_7)}{P(V_5 \wedge V_6)} \cdot \frac{P(V_6 \wedge V_8)}{P(V_6)} \end{aligned}$$

The initially assessed Bayesian network has now been transformed into a decomposable Bayesian network. The scheme for evidence propagation proposed by Spiegelhalter and Lauritzen operates on this decomposable Bayesian network. We emphasize that for a specific problem domain the transformation has to be performed only once: each consultation of the system proceeds from the obtained decomposable Bayesian network.

Recall that for making probabilistic statements concerning the statistical variables discerned in a problem domain we have to associate with a decomposable Bayesian network a method for computing probabilities of interest from it and a method for propagating evidence through it. As far as computing probabilities from a decomposable Bayesian network is concerned, it will be evident that any probability which involves only variables occurring in one and the same clique can simply be computed locally from the marginal probability function on that clique.

The method for evidence propagation is less straightforward. Suppose that evidence becomes available that the statistical variable V has adopted a certain value, say v. For ease of exposition, we assume that the variable V occurs in one clique of the decomposable graph only. Informally speaking, propagation of this evidence amounts to the following. The vertices and the cliques of the decomposable graph are ordered anew, this time starting with the instantiated vertex. The ordering of the cliques then is taken as the order in which the evidence is propagated through the cliques. For each subsequent clique, the updated marginal probability function is computed locally using the computation scheme shown below; we use

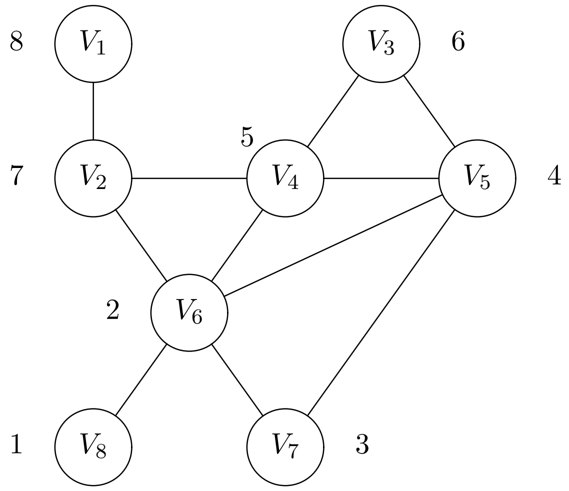


Figure 5.14: An ordering of the vertices starting with V_8.

P to denote the initially given probability function and P^* to denote the new probability function after updating. For the first clique in the ordering we simply compute:

$$P^*(C_{V(\mathrm{Cl}_1)}) = P(C_{V(\mathrm{Cl}_1)} \mid v)$$

For the remaining cliques, we compute the updated marginal probability function using:

$$\begin{aligned}
P^*(C_{V(\mathrm{Cl}_i)}) &= P(C_{V(\mathrm{Cl}_i)} \mid v) \\
&= P(C_{V(\mathrm{Cl}_i)\backslash S_i} \mid C_{S_i} \wedge v) \cdot P(C_{S_i} \mid v) \\
&= P(C_{V(\mathrm{Cl}_i)\backslash S_i} \mid C_{S_i}) \cdot P^*(C_{S_i}) \\
&= P(C_{V(\mathrm{Cl}_i)}) \cdot \frac{P^*(C_{S_i})}{P(C_{S_i})}
\end{aligned}$$

where S_i once more is the set of vertices Cl_i has in common with the lower numbered cliques. So, an updated marginal probability function is obtained by multiplying the 'old' marginal probability function with the quotient of the 'new' and the 'old' marginal probability function on the appropriate clique-intersection.

We look once more at our example.

EXAMPLE 5.22

Consider the decomposable graph from Figure 5.12 and its associated probability function once more. Suppose that we obtain the evidence that the variable V_8 has the value *true*. Using maximum cardinality search, we renumber the vertices of the graph starting with the vertex V_8. Figure 5.14 shows an example of such an ordering. From this new ordering of the vertices we obtain an ordering of the six cliques of the graph (once more, we identify a clique with its vertex set):

$$\begin{aligned}
\mathrm{Cl}_1 &= \{V_6, V_8\} \\
\mathrm{Cl}_2 &= \{V_5, V_6, V_7\} \\
\mathrm{Cl}_3 &= \{V_4, V_5, V_6\} \\
\mathrm{Cl}_4 &= \{V_3, V_4, V_5\}
\end{aligned}$$

$$\begin{aligned}\mathrm{Cl}_5 &= \{V_2, V_4, V_6\}\\ \mathrm{Cl}_6 &= \{V_1, V_2\}\end{aligned}$$

The impact of the evidence on the first clique is

$$P^*(V_6) = P(V_6 \mid v_8)$$

For the second clique we find:

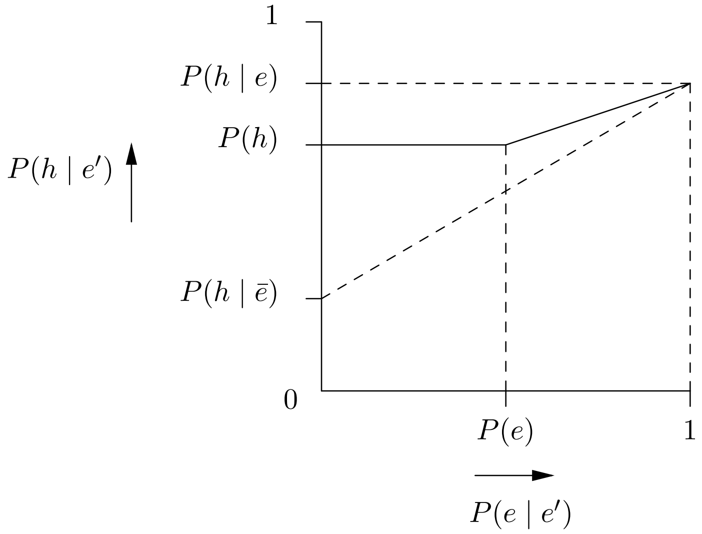
$$P^*(V_5 \wedge V_6 \wedge V_7) = P(V_5 \wedge V_6 \wedge V_7) \cdot \frac{P^*(V_6)}{P(V_6)}$$

For the remaining cliques we obtain similar results.

After the marginal probability functions have been updated locally, the instantiated vertex is removed from the graph, and the updated marginal probability functions are taken as the marginal probability functions on the cliques of the remaining graph. The process may now simply be repeated for a new piece of evidence.

Exercises

(5.1) The subjective Bayesian method uses a linear interpolation function as a combination function for propagating uncertain evidence. Recall that this interpolation function consists of two distinct linear functions, each defined on half of the domain of the combination function. Instead of the function employed in PROSPECTOR as discussed in Section 5.3.2, we could use for example the function shown in the figure below.



Describe the effect of applying the production rule **if** e **then** h **fi** on the prior probability of h in case this function is used as the combination function for uncertain evidence.

(5.2) Prove by means of counterexamples that the combination functions for composite evidence in the subjective Bayesian method are not correct when viewed from the perspective of probability theory.

(5.3) Write a PROLOG or LISP program implementing the subjective Bayesian method. You can depart from the program for top-down inference discussed in Chapter 3 (of Principles of Intelligent Systems).

(5.4) A particular rule-based system employs the certainty factor model for modelling the uncertainty that goes with the problem domain of concern. Let the following three production rules be given (only the names of the attributes in the conditions and conclusions are shown):

if b **or** c **then** $f_{0.3}$ **fi**
if f **and** g **then** $a_{0.8}$ **fi**
if d **or** e **then** $a_{0.2}$ **fi**

Furthermore, suppose that the attributes b, c, d, e, and g have been established with the certainty factors 0.2, 0.5, 0.3, 0.6, and 0.7, respectively. The attribute a is the goal attribute of top-down inference. Give the inference network resulting from top-down inference with these facts and production rules. Compute the certainty factor which results for the attribute a.

(5.5) Consider the following frame of discernment: $\Theta = \{a, b, c\}$. Let the basic probability assignment m be defined by $m(\{a\}) = 0.3$, $m(\{a, b\}) = 0.4$, $m(\{a, b, c\}) = 0.2$, $m(\{a, c\}) = 0.1$; the remaining basic probability numbers all equal 0. Compute $\text{Bel}(\{a, c\})$.

(5.6) Let Θ be a frame of discernment. Prove that for each $x \subseteq \Theta$ we have that $\text{Pl}(x) \geq \text{Bel}(x)$.

(5.7) Let $\Theta = \{a, b, c, d\}$ be a frame of discernment. Give an example of a basic probability assignment on Θ that defines a probability function on Θ at the same time.

(5.8) Consider the frame of discernment $\Theta = \{a, b, c\}$ and the following two basic probability assignments m_1 and en m_2:

$$m_1(x) = \begin{cases} 0.3 & \text{if } x = \Theta \\ 0.6 & \text{if } x = \{a, c\} \\ 0.1 & \text{if } x = \{b, c\} \\ 0 & \text{otherwise} \end{cases}$$

$$m_2(x) = \begin{cases} 0.8 & \text{if } x = \Theta \\ 0.2 & \text{if } x = \{b\} \\ 0 & \text{otherwise} \end{cases}$$

Construct the intersection tableau for the function $m_1 \oplus m_2$ using Dempster's rule of combination.

9. Consider the frame of discernment $\Theta = \{a, b, c\}$ and the following basic probability assignments m_1 and m_2:

$$m_1(x) = \begin{cases} 0.3 & \text{if } x = \Theta \\ 0.6 & \text{if } x = \{a, c\} \\ 0.1 & \text{if } x = \{a, b\} \\ 0 & \text{otherwise} \end{cases}$$

$$m_2(x) = \begin{cases} 0.8 & \text{if } x = \Theta \\ 0.2 & \text{if } x = \{a\} \\ 0 & \text{otherwise} \end{cases}$$

Why is it not necessary in this case to normalize? Compute the value of $\text{Bel}_1 \oplus \text{Bel}_2(\{a\})$.

(5.10) Consider the following medical information:

> Metastatic cancer is a possible cause of a brain tumor, and is also an explanation for increased total serum calcium. In turn, either of these could explain a patient falling into a coma. Severe headache is also possibly associated with a brain tumour.

Suppose that we use a Bayesian network to represent this information. Give the graphical part of the Bayesian network. Which probabilities have been associated with the graph?

(5.11) Consider the causal polytree from Figure 5.9 and an associated set of probabilities. Suppose that we apply the method of J.H. Kim and J. Pearl for evidence propagation. Try to find out how evidence spreads through the network if entered in one of the vertices.

(5.12) Consider the Bayesian network obtained in Exercise 10 once more. We transform this Bayesian network into a decomposable Bayesian network as described in Section 5.6.4.

(a) Give the resulting decomposable graph. Which cliques do you discern?

(b) Give the new representation of the originally given probability function.

(c) What happens if we obtain the evidence that a specific patient is suffering from severe headaches?