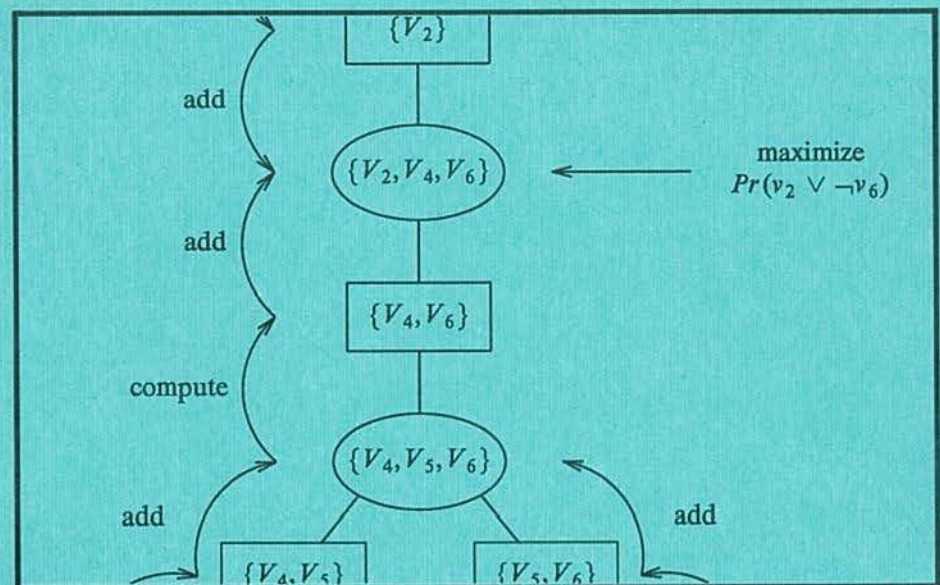


Probability-Based Models for Plausible Reasoning



Linda C. van der Gaag

Probability-Based Models for Plausible Reasoning

(Probabilistische Modellen voor het Redeneren met Onzekerheid)

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof.dr. S.K. Thoden van Velzen,
in het openbaar te verdedigen in de Aula der Universiteit
(Oude Lutherse kerk, ingang Singel 411, hoek Spui),
op woensdag 26 september 1990 te 13.30 uur

door

Linda Christina van der Gaag
geboren te Delft

Promotoren: prof.dr. J.A. Bergstra
prof.dr. R.D. Gill (Rijksuniversiteit Utrecht)

Faculteit: Wiskunde en Informatica

Acknowledgment

This thesis has been written while I was employed at the Centre for Mathematics and Computer Science. I like to thank the head of the Department of Software Technology, Jaco de Bakker, and the project leader of the Expert Systems Group, Peter Lucas, for letting me work on the subject of plausible reasoning all these years. Without the financial support and the freedom they have given me, the research reported would not have been possible.

Furthermore, I am grateful to Jan Bergstra and Richard Gill for being my thesis supervisors; it was Richard Gill who pointed out to me the paper by Steffen Lauritzen and David Spiegelhalter that inspired the main ideas of this thesis.

Most of all, however, I'd like to thank Steffen Lauritzen who invited me to Aalborg University in Denmark. My stay there has been extremely motivating and I have learned a great deal from him and Finn Verner Jensen. The many discussions we've had have turned out to be invaluable. Furthermore, I greatly appreciate Steffen Lauritzen's willingness to be a member of my reading committee.

Then, I also want to thank the other members of the reading committee, Peter van Emde Boas, Paul Klint and Marc Bezem, who provided comments on an earlier draft of this thesis.

I am particularly grateful to Peter Lucas, Hans Mulder, Roland Bol, Jan Willem Spee and Louis Kossen for providing comments, encouragement or both. To conclude, my special thanks go to Ay Ling Ong and Wouter Mettrop from the Library of the Centre for Mathematics and Computer Science for patiently gathering all the literature I thought I needed.

Notes

Section 2.1 has been published (under the title 'A conceptual model for inexact reasoning in rule-based systems') in the *International Journal of Approximate Reasoning*, vol. 3, no. 3, pp. 239 - 258, 1989.

An extended abstract of the Sections 4.2 and 4.3 has been excepted for publication (under the titel 'Computing probability intervals under independency constraints') in the *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, Cambridge, Massachusetts, 1990.

Contents

1. Introduction	1
2. Quasi-Probabilistic Models	9
2.1. A Syntactical Model for Rule-Based Plausible Reasoning	10
2.2. The Probabilistic Basis of the Certainty Factor Model	29
2.3. An Analysis of the Approximation Functions \overline{MB} and \overline{MD}	48
2.4. An Analysis of the Certainty Factor Functions CF , CF' and \overline{CF}	67
3. Belief Networks	81
3.1. Preliminaries: Graph Theory and Probability Theory	82
3.2. Knowledge Representation in a Belief Network	90
3.3. Evidence Propagation in a Belief Network	94
3.4. Evidence Propagation by Lauritzen and Spiegelhalter	95
4. Partially Quantified Belief Networks	117
4.1. Preliminaries: The Theory of Linear Programming	119
4.2. Partial Specification of a Joint Probability Distribution	121
4.3. Partial Quantification of a Belief Network	138
4.4. Processing Evidence	162
References	181
Samenvatting	189

Chapter 1

Introduction

During the past decade the interest in the results of artificial intelligence has been growing to an increasing extent. Especially the area of knowledge-based systems, one of the first areas of artificial intelligence to be commercially fruitful, has received a lot of attention. The phrase knowledge-based system is generally employed to denote systems in which some symbolic representation of human knowledge is incorporated and applied. Of these knowledge-based systems, the so-called *expert systems* have been the most successful at present. Expert systems are systems which are capable of offering solutions and advice concerning a specific real-life problem domain at a level comparable to that of experts in the same field (in this thesis, we do not consider an expert system an explicit model of human problem solving behaviour). The problems encountered in the fields for which expert systems are being developed are characterized by requiring considerable human expertise for their solution; examples of such domains are medical diagnosis, financial advice, product design etc. Expert systems therefore usually need large amounts of detailed expert knowledge to arrive at a performance comparable to that of human experts in the field.

When building expert systems it becomes evident that in many real-life domains expert knowledge is not precisely defined, but instead is of an imprecise nature. Yet, human experts typically are able to form judgments and take decisions from uncertain, incomplete and sometimes even contradictory information. In order to be useful in an environment in which only such imprecise knowledge is available, an expert system has to capture and exploit not only the highly-specialized expert knowledge, but the uncertainties that go with the represented pieces of knowledge as well. Researchers in artificial intelligence therefore have sought methods for representing uncertainty and have developed reasoning procedures for

manipulating uncertain knowledge. These efforts have grown into a major research topic in artificial intelligence called *inexact reasoning* or *plausible reasoning*.

Probability theory is one of the oldest mathematical theories concerning uncertainty. It is no wonder therefore that this formal theory was chosen as the first point of departure in the pioneering work on automated reasoning with uncertainty. During the 1960s several research efforts on probabilistic reasoning were undertaken, see for example [WARN61, GORR68, DOMB72]. The systems constructed in this period were primarily for (medical) diagnosis. These systems may be viewed as precursors of the diagnostic expert systems developed in the seventies. We take a closer look at the task of diagnosis for a better understanding of the problems the researchers were confronted with. For further information the reader is referred to [HORV88].

Let $H = \{h_1, \dots, h_n\}$, $n \geq 1$, be a set of n possible hypotheses, and let $E = \{e_1, \dots, e_m\}$, $m \geq 1$, 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, actually observed for the given case) or *false*. In real-life applications the relationships between the hypotheses and the evidence generally are uncertain. The diagnostic task then is to find a set of hypotheses $h \subseteq H$, called the *diagnosis*, which most likely accounts for the observed evidence. Now, let Pr be a probability distribution on the discerned sample space. If we have observed a set of pieces of evidence $e \subseteq E$, then, from a decision-theoretic, probabilistic point of view, we can simply compute the conditional probabilities $Pr(h | e)$ for each subset $h \subseteq H$, and select the set $h' \subseteq H$ with the highest probability. Since for real-life applications the conditional probabilities $Pr(e | h)$ often are easier to come by than the conditional probabilities $Pr(h | e)$, generally Bayes' Theorem is used for computing $Pr(h | e)$:

$$Pr(h | e) = \frac{Pr(e | h) Pr(h)}{Pr(e)}$$

It will be evident that the task of diagnosis in this form is computationally complex: because 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 as possible diagnoses the n singleton hypothesis sets $\{h_i\}$. So, we have to compute the probabilities (writing h_i instead of $\{h_i\}$)

$$Pr(h_i | e) = \frac{Pr(e | h_i) Pr(h_i)}{Pr(e)} = \frac{Pr(e | h_i) Pr(h_i)}{\sum_{k=1}^n Pr(e | h_k) Pr(h_k)}$$

for each $h_i \in H$ only. For a successful application of Bayes' Theorem in this form several conditional and prior probabilities are required. Note that in general the conditional probabilities $Pr(e|h_k)$ cannot be computed from their 'component' conditional probabilities $Pr(e_j|h_k)$, $e_j \in e$. Therefore, conditional probabilities $Pr(e|h_k)$ for every combination of pieces of evidence $e \subseteq E$ have to be available, that is, exponentially many probabilities have to be known. Since it is hardly likely that all these probabilities can be obtained, generally a second simplifying assumption is made: it is assumed that the pieces of evidence $e_j \in E$ are conditionally independent given any hypothesis $h_k \in H$. Under this assumption Bayes' Theorem reduces to

$$Pr(h_i | e_{j_1}, \dots, e_{j_p}) = \frac{Pr(e_{j_1} | h_i) \cdots Pr(e_{j_p} | h_i) Pr(h_i)}{\sum_{k=1}^n Pr(e_{j_1} | h_k) \cdots Pr(e_{j_p} | h_k) Pr(h_k)}$$

for the observed evidence $e = \{e_{j_1}, \dots, e_{j_p}\}$, $1 \leq p \leq m$. It will be evident that with the two assumptions mentioned above only mn conditional probabilities and $n - 1$ prior ones suffice for a successful use of Bayes' Theorem; for a more elaborate analysis, the reader is referred to [SZOL78].

The pioneering systems for probabilistic reasoning constructed in the 1960s 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 empirical data of several hundred sample cases. In spite of the (over-)simplifying assumptions underlying these systems they performed considerably well, see for example [DOMB74]. Nevertheless, interest in this decision-theoretic probabilistic approach to reasoning with uncertainty faded in the late 1960s and early 1970s, although it should be mentioned that this type of research is still pursued on a moderate scale, see for example [MALC86]. One of the reasons for this decline in interest is that the method is only feasible for highly restricted problem domains: for larger domains or domains in which the above-mentioned simplifying assumptions are seriously violated, the method inevitably becomes demanding, either computationally or from an assessment point of view.

At this stage, the first diagnostic expert systems began to emerge from the early artificial intelligence research efforts. These systems mostly used production rules, which in concept resembled logical implications, as a formalism for representing expert knowledge in a modular form and employed some heuristic reasoning method for applying the rules. These so-called *rule-based* expert systems exhibited an 'intelligent' reasoning behaviour as a consequence of their concentrating only on those hypotheses which were suggested by the evidence. In such systems, the production rules typically are used in selectively gathering evidence and heuristically pruning the search space of possible diagnoses. This pruning rendered the rule-based systems

capable of dealing with larger and complexer problem domains than the early decision-theoretic probabilistic systems were.

As we have mentioned before, to be useful for real-life applications the originally deterministic rule-based systems had to be extended with some notion of uncertainty. The two best-known systems developed in the 1970s which incorporated some method for dealing with uncertainty are the MYCIN system for assisting physicians in the diagnosis and treatment of bacterial infections, developed by B.G. Buchanan and E.H. Shortliffe, [BUCH84], and the PROSPECTOR system by R.O. Duda, P.E. Hart and others, [DUDA79], for aiding non-expert geologists in the identification of mineral deposits. For a probabilistic approach based on Bayes' Theorem such as employed in the systems of the 1960s a large number of conditional and prior probabilities was necessary, thus requiring enormous amounts of experimental data. In practice, the required data simply were not available; the necessary probabilities therefore could not be obtained from statistical analysis. In devising a probabilistic reasoning component to be incorporated in a rule-based system, the researchers therefore had to depend on subjective probabilities which had been assessed by human experts in the field.

The *subjectivist* or *personalist Bayesian* view to probability theory is opposed to the frequentist point of view generally adhered to from the beginning of this century. Although at present the subjectivist point of view is not (yet) popular with statisticians, it is becoming increasingly important in artificial intelligence research for the representation of uncertainty. Informally speaking, the subjectivist view to probability theory is the one 'the man in the street' takes. The central idea is that probabilistic statements may be made concerning any (potentially) verifiable proposition, independent of whether it is a statement concerning a repeatable experiment or not: a subjectivist views the probability of an event as a measure of a person's belief in the occurrence of the event, given the information that person has. The (philosophical) foundation for this approach to probability theory was established mainly by L.J. Savage, [SAVA54], and B. de Finetti, [FINE70]. From a subjectivist viewpoint, it is in principle possible for a domain expert to assess a probability for any proposition even if he knows little about it. Nevertheless, in practice an expert often is uncertain and uncomfortable about the probabilities he is providing. The difficulty of assessing probabilities is well-known as a result of research on human decision making and judgment under uncertainty, see for example [KAHN82]. In the present thesis, we do not discuss this issue any further; we merely build on the observation that domain experts generally are unable to fully specify a probability distribution on the problem domain, and furthermore, that the subjective assessments obtained from the experts are likely to be inconsistent.

To return again to rule-based expert systems, it will be evident that a probabilistic reasoning component to be integrated into such a system should be able to deal with a partial and often even an inconsistent specification of a probability distribution. In a rule-based context, an expert typically is asked to associate probabilities only with the production rules he has provided. We

have mentioned before that these production rules are used in pruning the search space of possible diagnoses; in this pruning process heuristical as well as probabilistic criteria are employed. It therefore becomes 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. To overcome these problems, in the 1970s several modifications of probability theory for efficient application in a rule-based environment were developed. These models offered computation rules which did not always accord with the axioms of probability theory but which rendered the models to some extent insensitive to partial specification and inconsistency of a probability distribution. None of these modifications, however, presents a theoretically well-founded solution to the above-mentioned problems: the models mostly have an ad hoc character. This observation inspired us to use the phrase *quasi-probabilistic models* to denote such models for dealing with uncertainty in rule-based expert systems. The two most well-known quasi-probabilistic models are the *certainty factor model*, [SHOR84], originally designed for dealing with uncertainty in the MYCIN system, and the *subjective Bayesian method*, [DUDA76], developed for the PROSPECTOR system we mentioned before. Especially the certainty factor model has since its introduction enjoyed widespread use in rule-based expert systems built after MYCIN. Even though the model is not well-founded from a mathematical point of view, in practice it seems to behave 'satisfactorily', see for example [SHOR84]. The relative success of the model can further be accounted for by its computational simplicity.

Since research on plausible reasoning has not yet yielded alternative models which are at the same time mathematically correct, computationally feasible and semantically clear, the early quasi-probabilistic models are still employed frequently in present-day rule-based expert systems. We feel that the frequent use of these models justifies an in-depth study. In Chapter 2 we introduce a conceptual model for plausible reasoning in a rule-based expert system showing the syntactical requirements an actual (quasi-probabilistic) model has to meet; this conceptual model is subsequently employed in studying the certainty factor model. The chapter discusses several results of an analysis of the probabilistic foundation of the certainty factor model; for example, it is shown that in the model rather strong independency assumptions are made implicitly. Even though it has been known since the late 1970s that the certainty factor model is mathematically flawed, most of the detailed results presented are new.

Although the quasi-probabilistic models on the one hand met with considerable success in the artificial intelligence community, they were criticized severely because of their ad hoc character on the other hand. The incorrectness of these models from a mathematical point of view and an analysis of the problems their developers were confronted with, even led to a world-wide discussion concerning the appropriateness of probability theory for handling uncertainty in a knowledge-based context. Two often-cited papers taking opposite positions are [CHEE85] and [ZADE86].

The adversaries of probability theory argue that it is not expressive enough to cope with the different kinds of uncertainty that are encountered in real-life situations and therefore have to be dealt with in expert systems. As a consequence several other (more or less) mathematical models for dealing with uncertainty have been proposed. A major trend in plausible reasoning has arisen from the claim that probability theory is not able to capture imprecision or vagueness, notions of uncertainty which are salient in natural language representations. The name of L.A. Zadeh is inseparable from this trend: he was the first to propose *fuzzy set theory* as the point of departure for the development of models which are able to cope with vague information. *Dempster-Shafer theory* is the other major trend in plausible reasoning. The theory has mainly been developed by G. Shafer, [SHAF76], out of earlier work by A.P. Dempster, [DEMP68]. It has evolved from the observation that probability theory is not able to discern between uncertainty and ignorance due to incompleteness of information. The two mentioned approaches are based on numerical assessments just like probability theory is. Some non-numerical methods for dealing with uncertainty have been proposed as well, see for example [COHE85,MCDE80].

The advocates of probability theory claim that it is provable that probability theory is the only correct way of dealing with uncertainty and that anything that in this context can be done with non-probabilistic techniques, can be done equally well using a probability-based method, see for example [HORV86,CHEE88]; for this claim often an argument by R.T. Cox is cited, [COX79], which states a simple set of intuitive properties a measure of uncertainty has to satisfy and subsequently shows that the (standard) axioms of probability theory follow. In addition it is often pointed out that probability theory is a mathematically well-founded theory having a long and outstanding tradition of research and experience. Concerning this point, we feel that giving up on probability theory for plausible reasoning just on account of the mathematically disappointing quasi-probabilistic models indeed is giving up too easily. In this thesis, we do not enter into the heated debate concerning the appropriateness of probability theory for handling uncertainty in a knowledge-based setting. For a wide range of diverging opinions, the reader is referred to [KANA86,LEMM88,KANA89]; also [LAUR88a] and [CHEE88] with their respective discussions are worth reading.

Although the above-mentioned discussion has not in the least subdued, in the mid-eighties a new trend in probabilistic reasoning in an artificial intelligence context became discernable: several models have been proposed departing from a graphical representation of a problem domain, see for example [SHAC86,PEAR88,SPIE86b]. Hereafter such a graphical representation will be called a *belief network*. Informally speaking, a belief network is a map of the statistical variables discerned in the problem domain and the probabilistic independency relationships holding between them. These interrelationships are quantified by means of 'local' conditional probabilities together defining a joint probability distribution on the variables. Associated with a belief network are a method for propagating the impact of evidence and

a method for computing the probabilities of interest for diagnosis. These methods employ the belief network as an architecture for performing certain probabilistic computations which involve small subsets of variables only. In Chapter 3 we present a formal description of the notion of a belief network, which in the relevant literature often is introduced only informally. We further discuss the work on belief networks by S.L. Lauritzen and D.J. Spiegelhalter, [LAUR88a]. In this thesis, we do not address the problem of constructing a belief network for a given domain; for information on this subject, the reader is referred to [HENR89, REBA89].

If we compare the work on belief networks with the pioneering work on probabilistic reasoning from the 1960s on the one hand and the quasi-probabilistic models from the 1970s on the other hand, then the belief network models in a sense are 'closer' to the former than to the latter. The belief network models again require a total and consistent specification of a joint probability distribution on the variables discerned in the problem domain. However, instead of making the oversimplifying general assumption of conditional independence, distinguished independency relationships between the variables are represented explicitly in the belief network. As has been mentioned before, the belief network is further exploited for restricting the necessary probabilistic computations to local ones on small sets of variables. However, only if many independency relationships hold between the variables is the method feasible. Just like the early decision-theoretic systems of the 1960s the systems employing a belief network model therefore are able to deal with rather restricted problem domains only. For larger domains again the problem arises that for application of a belief network model a large number of conditional and prior probabilities is necessary, with all the unpleasant consequences we have encountered before in the quasi-probabilistic models; for example, if the probabilities required for exact probabilistic reasoning cannot be obtained from statistical analysis, we once more have to rely on subjective probabilities assessed by human experts in the field. In their present form, the belief network models are not capable of dealing with a partial specification of a joint probability distribution, nor with an inconsistent one. To overcome part of this problem, we propose in Chapter 4 a method for computing upper and lower bounds on probabilities of interest from a *partially quantified belief network*; informally speaking, a partially quantified belief network is a kind of belief network in which all independency relationships between the statistical variables discerned are known or are assessed by a domain expert, but in which the initially given probabilities do not give rise to a unique joint probability distribution on the variables. The general idea of our method is to take the subjective probabilities assessed by the domain experts as defining constraints on a yet unknown joint probability distribution. Unfortunately, we have not been able to supplement our method for computing probability intervals with a method for propagating the impact of evidence. The method for computing lower and upper bounds on probabilities, however, may be used as a help in assessing the probabilities required for a fully specified belief network.

Although we are aware that many problems still remain to be solved, we hope to bring the belief network models with our method one step closer to exploitation in expert systems for real-life applications.

Chapter 2

Quasi-Probabilistic Models

In the early years of artificial intelligence research on plausible reasoning, efforts were concentrated on the application of probability theory in rule-based expert systems. As we have discussed in Chapter 1, it soon became evident that this mathematical theory could not be applied in such a context in a straightforward manner. The researchers were confronted with several problems, such as:

- It often is not possible to obtain a fully specified probability function on the domain of concern: only a few probabilities are known or can be estimated by an expert in the field, that is, often only a partial specification of a probability function is available.
- In case an expert has assessed many of the required probabilities, the thus obtained partial specification of a probability function is likely to be inconsistent in the sense that there is not an 'underlying' actual probability function.
- Probability theory does not provide *explicit computation rules* for computing probabilities from a partial specification of a probability function for all (intermediate) results derived from applying the production rules during an actual consultation of the system.

For a short period of time, research centered around the development of modifications of probability theory that should overcome these problems and that could be applied efficiently in a rule-based environment. The quasi-probabilistic models which resulted from these efforts are the topic of this chapter.

A model for reasoning with uncertainty has to exhibit a number of characteristics before it is applicable in a knowledge-based system using

production rules for knowledge representation. In Section 2.1 we introduce a conceptual model for plausible reasoning in a rule-based top-down reasoning expert system, showing the syntactical requirements an actual model has to meet. This syntactical model is used to examine the well-known *certainty factor model*. We show that our model has enabled us to elucidate some syntactical inadequacies in the original notation used by its developers. From these observations we arrive at a syntactically correct reformulation of the model without affecting its intended meaning. In the three subsequent sections we use this reformulation as a point of departure for an analysis of the model given its probabilistic foundation. We show for example that in the model several rather strong assumptions are implicitly made.

2.1. A SYNTACTICAL MODEL FOR RULE-BASED PLAUSIBLE REASONING

In this section, we show that a model for rule-based reasoning with uncertainty has to meet a number of syntactical requirements. A syntactical model explicitly stating these requirements is proposed, providing us with a conceptual framework for examining and comparing quasi-probabilistic models.

Although we assume that the reader is acquainted with production rules and top-down inference, we start with a brief description of these notions in order to introduce some terminology. For a more elaborate introduction, the reader is referred to [BUCH83, LUCA90]. In a rule-based top-down reasoning expert system applying the certainty factor model for the manipulation of uncertainty, three major components are discerned:

- (1) *Production rules* and associated *certainty factors*. Basically, an expert in the domain in which the expert system is to be used, models his knowledge of the field in a set of production rules of the form $e \rightarrow h$, which closely resemble logical implications. The left-hand side e of a production rule is a positive Boolean combination of conditions, that is, e does not contain any negation symbols. Without loss of generality we assume that e is a conjunction of disjunctions of conditions. Throughout this chapter, e as well as its constituting parts will be called (*pieces of*) *evidence*. In general, the right-hand side h of a production rule is a conjunction of conclusions. In the sequel, we restrict ourselves to single-conclusion production rules; note that this restriction is not an essential one from a logical point of view. From now on, a conclusion will be called a *hypothesis*. A production rule has the following meaning: if evidence e has been observed, then the hypothesis h is true.

An expert associates with the hypothesis h of each production rule $e \rightarrow h$ a (real) number $CF(h, e, e \rightarrow h)$, quantifying the degree to which the observation of evidence e confirms the hypothesis h . The values $CF(x, y, z)$ of the (partial) function CF are called *certainty factors*; $CF(x, y, z)$ should be read as ‘the certainty factor of x , given y and the derivation z of x from y ’. In the sequel we will use the more suggestive notation $CF(x \dashv y, z)$. (In [SHOR84], the developers of the certainty

factor model, E.H. Shortliffe and B.G. Buchanan, use for certainty factors the two-argument notation $CF(h,e)$; as will be discussed shortly, it is necessary to introduce the notion of a derivation in the notational convention.) Certainty factors range from -1 to $+1$. A certainty factor greater than zero is associated with a hypothesis h given some evidence e if the hypothesis is confirmed to some degree by the observation of this evidence; the certainty factor $+1$ indicates that the occurrence of evidence e completely proves the hypothesis h . A negative certainty factor is suggested if the observation of evidence e disconfirms the hypothesis h . A certainty factor equal to zero is suggested by the expert if the observation of evidence e does not influence the confidence in the hypothesis h .

- (2) *User-supplied data* and associated certainty factors. During a consultation of the expert system, the user is asked to supply actual case data. We assume that the user is not allowed to retract or modify earlier provided data; furthermore, we assume that in supplying data the user adheres to the same 'world view' throughout the consultation. The user attaches a certainty factor $CF(e \leftarrow u, u \rightarrow e)$ to every piece of evidence e he supplies the system with. In order to be able to treat production rules and user-supplied data uniformly, we assume the set of production rules supplied by the expert to be augmented with a set of fictitious production rules $u \rightarrow e$ for every piece of user-supplied evidence e , where u is taken to represent the user's de facto knowledge.
- (3) A (*top-down*) *inference engine* and a (*bottom-up*) *scheme for propagating uncertainty*. Top-down inference is a goal-directed reasoning technique in which the production rules are applied exhaustively to prove one or more goal hypotheses. A production rule is said to *succeed* if each of its conditions is fulfilled; otherwise, the rule is said to *fail*. Due to the application of production rules, several intermediate hypotheses will be confirmed to some degree during the inference process. The certainty factor to be associated with such an intermediate hypothesis h is calculated from the certainty factors associated with the production rules that were used in deriving h . For the purpose of thus propagating uncertainty, several functions for combining certainty factors are defined.

For those readers who are already familiar with the certainty factor model it is noted that in the sequel we abstract from several pragmatical issues involved in the model such as for example the discontinuity of the evaluation of the left-hand side of a production rule (that is, the 0.2 threshold).

2.1.1. Rule-Based Derivations and Derivation Trees

In the foregoing we have discussed the basic notions of the certainty factor model in an informal manner. In this section some formal definitions are provided.

DEFINITION 2.1. Let \mathcal{A} denote a set of atomic propositions. Let \mathcal{E} denote the set of conjunctions of disjunctions of elements of \mathcal{A} , that is, \mathcal{E} contains elements of the form $\bigwedge_{i=1}^n (\bigvee_{j=1}^{m_i} a_{ij})$, $a_{ij} \in \mathcal{A}$, $n, m_i \geq 1$, $i = 1, \dots, n$.

A hypothesis is an element $h \in \mathcal{A}$. A piece of evidence is an element $e \in \mathcal{E}$. Let u be a fixed element of \mathcal{A} representing the user's de facto knowledge.

A production rule is an expression $e \rightarrow h$ where e is a piece of evidence and h is a hypothesis.

In the sequel, we will omit parenthesis from elements of \mathcal{E} as long as ambiguity cannot occur.

We have informally introduced the certainty factor function having three arguments; recall that the 'third' argument of a certainty factor $CF(x \dashv y, z)$ represents a derivation of the hypothesis x from y with respect to a given set of production rules. The notion of a derivation is defined formally in the following definition.

DEFINITION 2.2. Let \mathcal{E} be defined as above. Furthermore, let \mathcal{P} be a finite, non-empty set of production rules. A derivation D^{ij} of j from i , $i, j \in \mathcal{E}$, with respect to \mathcal{P} is defined recursively as follows:

- (1) $e \rightarrow h$ is a derivation of h from e with respect to \mathcal{P} if $e \rightarrow h \in \mathcal{P}$.
- (2) If $D^{u,e}$ is a derivation of e from u with respect to \mathcal{P} and $e \rightarrow h$ is a derivation of h from e with respect to \mathcal{P} , then $(D^{u,e}) \circ (e \rightarrow h)$ is a derivation of h from u with respect to \mathcal{P} ; $(D^{u,e}) \circ (e \rightarrow h)$ is called the sequential composition of the derivations $D^{u,e}$ and $e \rightarrow h$.
- (3) If D^{u,e_1} is a derivation of e_1 from u with respect to \mathcal{P} and D^{u,e_2} is a derivation of e_2 from u with respect to \mathcal{P} , then $(D^{u,e_1}) \& (D^{u,e_2})$ is a derivation of $e_1 \wedge e_2$ from u with respect to \mathcal{P} ; $(D^{u,e_1}) \& (D^{u,e_2})$ is called the conjunction of the derivations D^{u,e_1} and D^{u,e_2} .
- (4) If D^{u,e_1} is a derivation of e_1 from u with respect to \mathcal{P} and D^{u,e_2} is a derivation of e_2 from u with respect to \mathcal{P} and if $e_1 \vee e_2 \in \mathcal{E}$, then $(D^{u,e_1}) \mid (D^{u,e_2})$ is a derivation of $e_1 \vee e_2$ from u with respect to \mathcal{P} ; $(D^{u,e_1}) \mid (D^{u,e_2})$ is called the disjunction of the derivations D^{u,e_1} and D^{u,e_2} .
- (5) If $D_1^{u,h}$ and $D_2^{u,h}$ are derivations of h from u with respect to \mathcal{P} , then $(D_1^{u,h}) \parallel (D_2^{u,h})$ is a derivation of h from u with respect to \mathcal{P} ; $(D_1^{u,h}) \parallel (D_2^{u,h})$ is called the parallel composition of the derivations $D_1^{u,h}$ and $D_2^{u,h}$.

The set of all derivations with respect to \mathcal{P} will be denoted by \mathcal{D} .

We remark that the notion of sequential composition is defined asymmetrically. Although a symmetrical definition $(D^{u,e}) \circ (D^{e,h})$ would be more appealing, it does not reflect the notion of a derivation as it occurs in

top-down inference. In the sequel, we will omit parentheses from elements of \mathcal{D} as long as ambiguity cannot occur.

EXAMPLE 2.3. Let \mathcal{P} be the set consisting of the following production rules:

$$\begin{array}{ll} d \wedge f \rightarrow b & u \rightarrow a \\ a \rightarrow d & u \rightarrow b \\ b \rightarrow i & u \rightarrow f \end{array}$$

Then, $D^{u,d} = (u \rightarrow a) \circ (a \rightarrow d)$ is a derivation of d from u , and $D^{u,i} = (((u \rightarrow b) \parallel (((u \rightarrow a) \circ (a \rightarrow d)) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b))) \circ (b \rightarrow i)$ is a derivation of i from u . ■

We conclude this subsection by introducing a graphical representation for derivations. A graphical representation of a derivation is called a *derivation tree*. Since our notion of a derivation tree is rather straightforward, we will confine ourselves to loosely introducing the basic building blocks for derivation trees.

Let $\rho(D)$ denote the graphical representation of the derivation D . We define

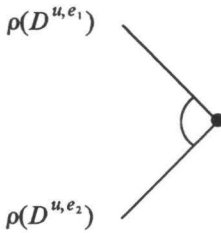
- (1) for the representation of a production rule $u \rightarrow h$,

$$\rho(u \rightarrow h) = \quad u \quad \longrightarrow \quad h$$

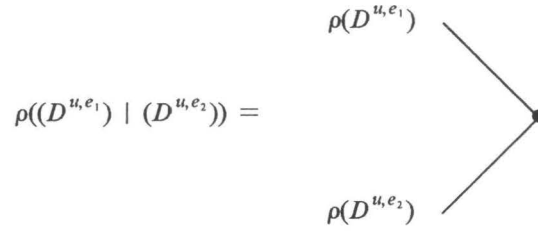
- (2) for the representation of the sequential composition of two derivations,

$$\rho((D^{u,e} \circ (e \rightarrow h)) = \quad \rho(D^{u,e}) \quad \longrightarrow \quad h$$

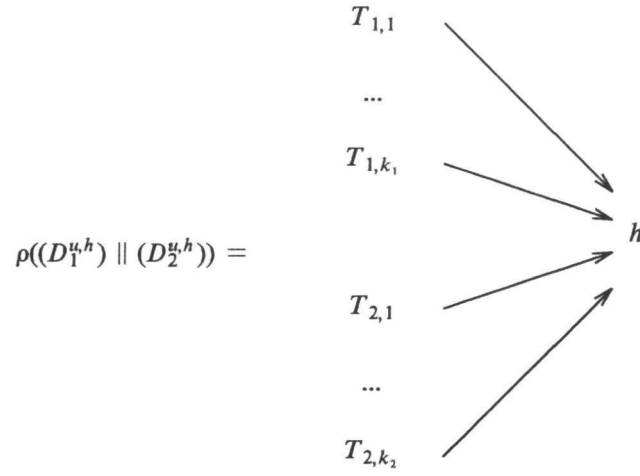
- (3) for the representation of the conjunction of two derivations,

$$\rho((D^{u,e_1} \& (D^{u,e_2})) =$$


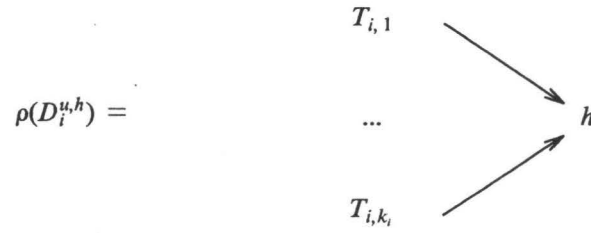
- (4) for the representation of the disjunction of two derivations,



- (5) and for the representation of the parallel composition of two different derivations of the same hypothesis h ,



where $k_1, k_2 \geq 1$, and for $i = 1, 2$,



that is, we simply join the two derivation trees $\rho(D_1^{u,h})$ and $\rho(D_2^{u,h})$.

A derivation tree for a derivation D is built by recursively using these basic representations.

EXAMPLE 2.4. Consider the set of production rules from Example 2.3 once more. The derivation tree $\rho(D^{u,b})$ of the derivation $D^{u,b} = ((u \rightarrow b) \parallel (((u \rightarrow a) \circ (a \rightarrow d)) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b)))$ is shown in Figure 2.1. ■

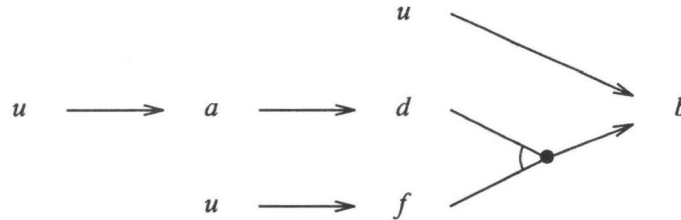


FIGURE 2.1. A derivation tree.

2.1.2. Rule-Based Inference and Inference Networks

In this subsection we show with the help of an example that an expert system with a fixed set of production rules applying the MYCIN top-down reasoning strategy determines a derivation in the set of all derivations with respect to this set of production rules. We assume that a backward-chaining strategy is used, that is, the production rules are applied in the order in which they have been specified; equally, the conditions in a production rule are evaluated in the specified order. Furthermore, we assume that the user is asked to confirm or disconfirm to some degree each piece of evidence that cannot be derived from the production rules.

EXAMPLE 2.5. Consider the set of production rules consisting of the following six elements:

$$\begin{array}{ll} e \rightarrow h & f \vee g \rightarrow h \\ a \wedge (b \vee c) \rightarrow h & a \rightarrow d \\ d \wedge f \rightarrow b & b \rightarrow i \end{array}$$

Note that the set of rules has not yet been supplemented with the fictitious production rules representing the user-supplied evidence. We suppose for the moment that h is the goal hypothesis of the consultation. First, the rule $e \rightarrow h$ is selected for application; e now becomes the next goal hypothesis to be

confirmed. As there are no production rules concluding on e , the user is asked to confirm or disconfirm e . We assume that he disconfirms e , in which case the production rule $e \rightarrow h$ fails. Subsequently, the user is asked to confirm or disconfirm a . When a is confirmed, f will be asked, and so on. We assume that a , c , f and g are confirmed by the user. So, the production rules $a \wedge (b \vee c) \rightarrow h$, $d \wedge f \rightarrow b$, $f \vee g \rightarrow h$ and $a \rightarrow d$ succeed. Note that the production rule $b \rightarrow i$ is not used in the derivation of h . ■

A top-down inference process as discussed in the preceding example may be depicted in a so-called *inference network*. An inference network is built from the representations of those production rules that actually succeeded during the inference process. In depicting inference networks we use building blocks similar to the ones introduced in Section 2.1.1 for the graphical representation of derivations.

EXAMPLE 2.6. Consider the top-down inference process described in Example 2.5 once more. The inference network corresponding with this inference process is shown in Figure 2.2. Note that the figure does not show the production rule $e \rightarrow h$ which has failed during the inference process nor the production rule $b \rightarrow i$ which has not even been selected for application. Furthermore, the figure does not depict as yet the user-supplied evidence. ■

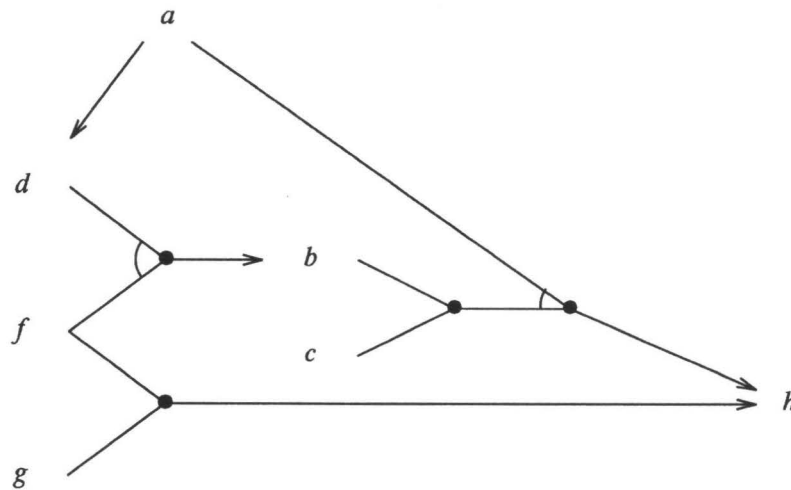


FIGURE 2.2. An inference network.

An inference network is subsequently extended with (fictitious) production rules $u \rightarrow e$ for each piece of user-supplied evidence e relevant to the production rules that actually succeeded during the consultation of the system in deriving one or more goal hypotheses. Recall that u represents the user's de facto knowledge.

EXAMPLE 2.7. The inference network of Figure 2.2 is extended with the production rules $u \rightarrow a$, $u \rightarrow c$, $u \rightarrow f$ and $u \rightarrow g$ since the user confirmed a , c , f and g . The thus extended inference network is depicted in Figure 2.3.

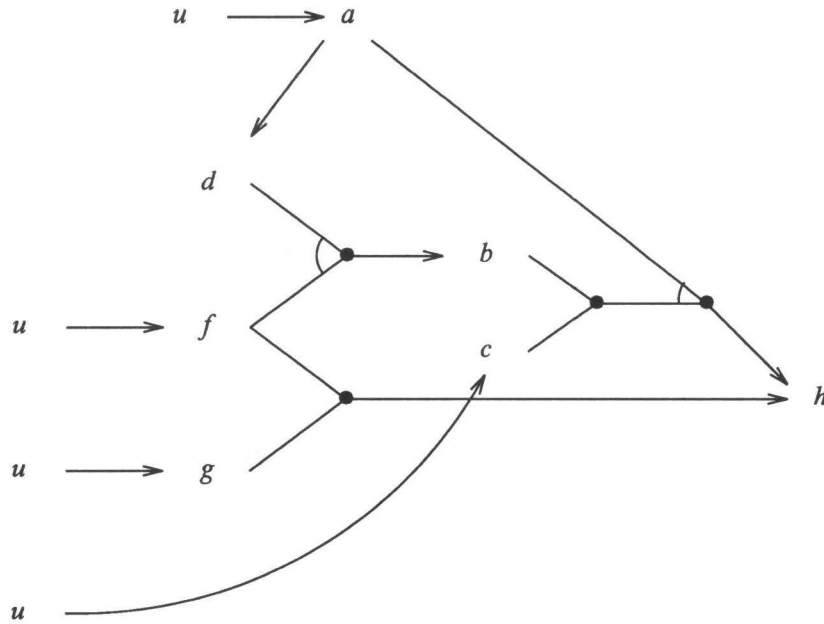


FIGURE 2.3. The extended inference network.

Now, consider the production rule $a \wedge (b \vee c) \rightarrow h$ once more. Up till now we have assumed that a and c were both confirmed by the user and that b was derived by applying some of the other rules. The reader can easily verify that the mentioned rule also succeeds in the case in which b has been derived like before, and in which the user has confirmed a but has *disconfirmed* c . In this case, the inference network will be exactly the same as the one shown in Figure 2.2. In the sequel, we like to treat the different instances of success of a production rule specifying a disjunction in its left-hand side uniformly. Therefore, in all such instances, an inference network comprising the rule is extended in the same way. So, even though the user has supplied negative

information on c in our example, the inference network after extension is the same as the one shown in Figure 2.3. Notice the difference between the treatment of the disconfirmation of c and the disconfirmation of e which led to the failure of the production rule $e \rightarrow h$. ■

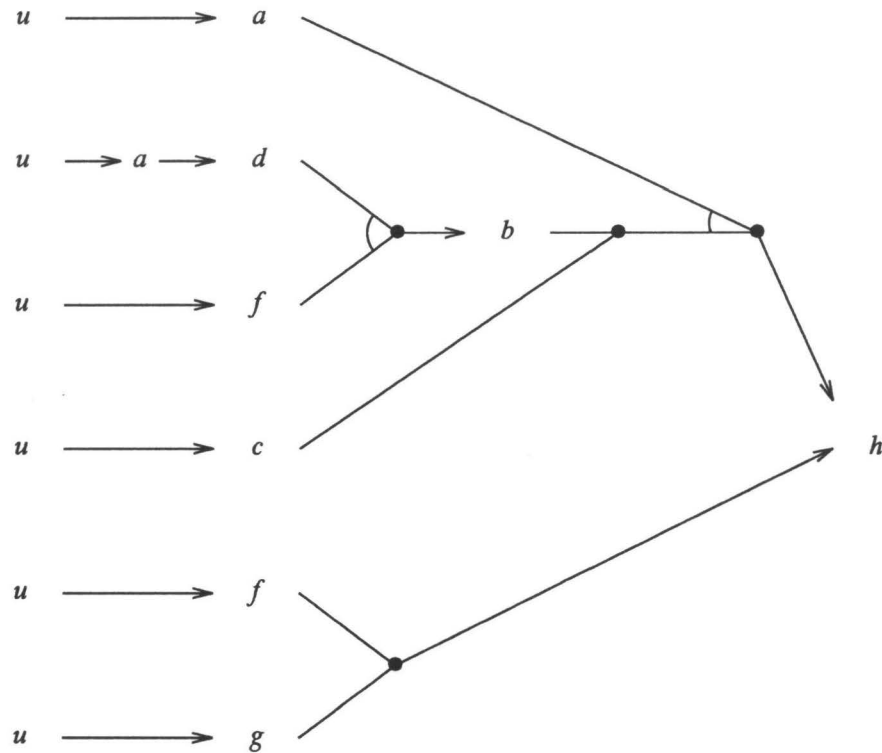


FIGURE 2.4. The transformed inference network.

It is noted that when using the MYCIN top-down reasoning strategy, each production rule may be applied at most once during an inference process. Furthermore, the inference networks composed of only those production rules that actually succeeded during a consultation are guaranteed to be acyclic since the reasoning mechanism prevents cyclic reasoning chains. From the latter observation we have that each extended inference network can be transformed in such a way that from each vertex representing an element from \mathcal{A} either one arrow of type $\rightarrow \bullet$ or one arrow of type \rightarrow departs; this may be achieved by duplicating certain vertices and arrows. So, an inference network may be transformed into a tree-like structure.

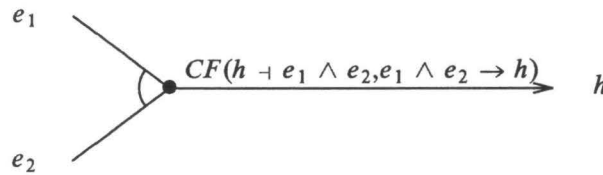
EXAMPLE 2.8. Figure 2.4 shows the inference network resulting from the transformation of the extended inference network depicted in Figure 2.3. Notice the duplication of the vertices a and f . ■

A transformed inference network now equals a derivation tree representing an element of the set of all derivations with respect to the given set of production rules.

2.1.3. Modelling the Propagation of Uncertainty

In the preceding subsections we have shown that an instance of a rule-based top-down reasoning inference process can be represented graphically as an inference network which after extension and transformation corresponds with a derivation tree. In this section, we use such extended and transformed tree-like inference networks to model the propagation of uncertainty during an inference process. Henceforth, the phrase 'inference network' will be used to denote an extended and transformed inference network.

Recall that an expert has attached a certainty factor $CF(h \leftarrow e, e \rightarrow h)$ to the conclusion h of each production rule $e \rightarrow h$, and that the user has associated a certainty factor $CF(e \leftarrow u, u \rightarrow e)$ with the conclusion e of each fictitious production rule $u \rightarrow e$ representing the fact that he has supplied the system with the factual information e . In an inference network, a certainty factor assigned to the hypothesis of a production rule will be attached to the arrow in the representation of the rule. So, if an expert has associated the certainty factor $CF(h \leftarrow e_1 \wedge e_2, e_1 \wedge e_2 \rightarrow h)$ with the hypothesis h in the rule $e_1 \wedge e_2 \rightarrow h$, this is represented as shown below:



The aim of the certainty factor model is to calculate the certainty factor $CF(h \leftarrow u, D^{u,h})$ for each goal hypothesis h where $D^{u,h}$ is the derivation of h from u with respect to a fixed set of production rules which are applied exhaustively in a top-down reasoning fashion; it will be evident that this certainty factor is dependent upon the certainty factors attached to the arrows in the inference network as well as on the structure of the inference network itself.

The way the certainty factor $CF(h \vdash u, D^{u,h})$ for a goal hypothesis h is calculated from other certainty factors is modelled with the help of the inference network. We define a number of basic *compression steps* that are used to compress an inference network in a finite number of steps to

$$u \xrightarrow{CF(h \vdash u, D^{u,h})} h$$

for each goal hypothesis h . As we will describe shortly, in each compression step the number of arrows (and certainty factors) in the network decreases. The certainty factors that disappear in a compression step are combined into a new certainty factor. For that purpose a so-called *combination function* is associated with each compression step. There are four basic compression steps:

(1) The inference network

$$u \xrightarrow{CF(e \vdash u, D^{u,e})} e \xrightarrow{CF(h \vdash e, e \rightarrow h)} h$$

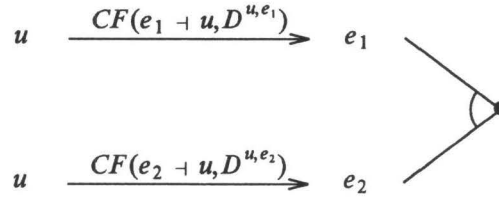
representing the sequential composition of the derivations $D^{u,e}$ and $e \rightarrow h$ is compressed yielding:

$$u \xrightarrow{CF(h \vdash u, (D^{u,e}) \circ (e \rightarrow h))} h$$

With this compression step, a combination function f_{\circ} is associated such that

$$CF(h \vdash u, (D^{u,e}) \circ (e \rightarrow h)) = f_{\circ}(CF(e \vdash u, D^{u,e}), CF(h \vdash e, e \rightarrow h))$$

(2) The inference network



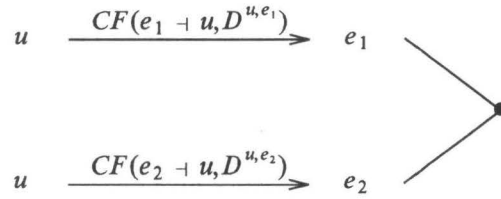
representing the conjunction of the derivations D^{u,e_1} and D^{u,e_2} is compressed yielding:

$$u \xrightarrow{CF(e_1 \wedge e_2 \dashv u, (D^{u,e_1}) \& (D^{u,e_2}))} e_1 \wedge e_2$$

With this compression step, a combination function $f_{\&}$ is associated such that

$$CF(e_1 \wedge e_2 \dashv u, (D^{u,e_1}) \& (D^{u,e_2})) = f_{\&}(CF(e_1 \dashv u, D^{u,e_1}), CF(e_2 \dashv u, D^{u,e_2}))$$

(3) The inference network



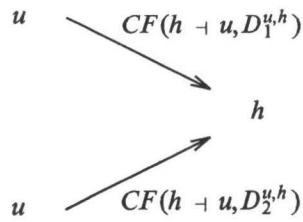
representing the disjunction of the derivations D^{u,e_1} and D^{u,e_2} is compressed yielding:

$$u \xrightarrow{CF(e_1 \vee e_2 \dashv u, (D^{u,e_1}) \mid (D^{u,e_2}))} e_1 \vee e_2$$

With this compression step, a combination function f_{\mid} is associated such that

$$CF(e_1 \vee e_2 \dashv u, (D^{u,e_1}) \mid (D^{u,e_2})) = f_{\mid}(CF(e_1 \dashv u, D^{u,e_1}), CF(e_2 \dashv u, D^{u,e_2}))$$

(4) The inference network



representing the parallel composition of the derivations $D_1^{u,h}$ and $D_2^{u,h}$ is compressed yielding:

$$u \xrightarrow{CF(h \dashv u, (D_1^{u,h} \parallel D_2^{u,h}))} h$$

With this compression step, a combination function f_{\parallel} is associated such that

$$CF(h \dashv u, (D_1^{u,h} \parallel D_2^{u,h})) = f_{\parallel}(CF(h \dashv u, D_1^{u,h}), CF(h \dashv u, D_2^{u,h}))$$

Since application of each of these four compression steps reduces the number of arrows (and certainty factors) in an inference network, termination of the compression is guaranteed.

EXAMPLE 2.9. The inference network of Figure 2.4 is compressed to

$$u \xrightarrow{CF(h \dashv u, D^{u,h})} h$$

where $D^{u,h} = (((u \rightarrow a) \& (((((u \rightarrow a) \circ (a \rightarrow d)) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b)) \mid \mid (u \rightarrow c))) \circ (a \wedge (b \vee c) \rightarrow h)) \parallel (((u \rightarrow f) \mid (u \rightarrow g)) \circ (f \vee g \rightarrow h))$. ■

In the foregoing, we have introduced the basic compression steps and their associated combination functions relative to inference networks that had been extended and transformed. Recall that we transformed an inference network after extension to arrive at a tree-like structure equalling a derivation tree representing an element of the set of all derivations with respect to the given set of production rules. In actually using the combination functions in implementations of the model, however, certainty factors are only computed once; so in practice, the extended yet not transformed inference network is employed.

2.1.4. Some Desirable Properties of the Combination Functions

In the preceding subsection we have modelled the propagation of uncertainty during an inference process by means of compression of the corresponding inference network. We have defined four basic compression steps and have introduced combination functions corresponding with these compression steps. In this section we discuss these combination functions and indicate some desirable properties for each of them.

Recall that the certainty factor $CF(h \dashv e, e \rightarrow h)$ quantifies the degree to which the occurrence of evidence e confirms the hypothesis h . However, the

truth of a piece of evidence e (that is, whether or not e has actually been observed) may not always be determined with absolute certainty: with every piece of evidence e supplied by the user, a certainty factor is associated not necessarily equal to +1. Furthermore, when using production rules intermediate hypotheses are confirmed to some degree and may in turn be used as evidence in other production rules concluding on new hypotheses. The basic compression step (1) describing the composition of derivations, and its associated combination function f_0 deal with this situation. From now on, we will call the function f_0 the *combination function for (propagating) uncertain evidence*.

The evidence e in a production rule $e \rightarrow h$ may be an intermediate hypothesis which has been confirmed to some degree. If a certainty factor $CF(e \rightarrow u, D^{u,e})$ for the evidence e given some derivation of e from u is known, the combination function for propagating uncertain evidence can handle this situation. However, the evidence e in a rule $e \rightarrow h$ may be a conjunction of disjunctions of pieces of evidence. In order to be able to apply the combination function f_0 for uncertain evidence, the certainty factor $CF(e \rightarrow u, D^{u,e})$ of the positive Boolean combination e has to be computed from the certainty factors of its constituting parts. The basic compression steps (2) and (3) dealing with the conjunction and disjunction of derivations, and their associated combination functions $f_\&$ and $f_!$, respectively, refer to this situation. From now on, the function $f_\&$ will be called the *combination function for a conjunction of hypotheses*, and the function $f_!$ the *combination function for a disjunction of hypotheses*. When referring to the two functions together, we will call them the *combination functions for composite hypotheses*.

From a mathematical point of view, it is desirable that application of the combination functions for composite hypotheses render the resulting certainty factor for a conjunction of disjunctions of pieces of evidence independent of the order in which the constituting parts of each of the disjunctions and the order in which the constituting parts of each of the conjunctions are specified. For example, the production rules $e_1 \wedge e_2 \rightarrow h$ and $e_2 \wedge e_1 \rightarrow h$ should yield the same result. Furthermore, the certainty factor of a positive Boolean combination of pieces of evidence has to be independent of the way in which the constituting parts of each of the disjunctions and of the way in which the constituting parts of the conjunction are taken together to be combined. Therefore, the combination functions for composite hypotheses $f_\&$ and $f_!$ have to respect the property of commutativity

$$f_\&(x,y) = f_\&(y,x), \text{ and}$$

$$f_!(x,y) = f_!(y,x)$$

for all certainty factors x and y , and the property of associativity

$$f_\&(f_\&(x,y),z) = f_\&(x,f_\&(y,z)), \text{ and}$$

$$f_!(f_!(x,y),z) = f_!(x,f_!(y,z))$$

for all certainty factors x , y and z . Furthermore, specifying a hypothesis more than once in a conjunction or disjunction should not influence the resulting certainty factor. So, the combination functions for composite hypotheses have to respect the property of idempotency

$$f_{\&}(x, x) = x, \text{ and}$$

$$f_{\mid}(x, x) = x$$

for all certainty factors x . Note that assuming the mathematical properties of commutativity, associativity and idempotency in an explicit model of human problem solving behaviour under uncertainty might not be realistic from a psychological point of view.

When different successful production rules $e_i \rightarrow h$ (that is, rules with different left-hand sides e_i) conclude on the same hypothesis h , a certainty factor $CF(h \dashv u, (D^{u, e_i}) \circ (e_i \rightarrow h))$ is derived from the application of each of these production rules. The net certainty factor for h is dependent upon each of these partial certainty factors. The basic compression step (4) describing the parallel composition of derivations, and its associated combination function f_{\parallel} deal with such co-concluding production rules. From now on, we will call the function f_{\parallel} the *combination function for (combining the results of) co-concluding production rules*.

Again, it is desirable that application of the function f_{\parallel} renders the resulting certainty factor for a hypothesis h independent of the order in which the different production rules concluding on h are applied. Furthermore, it is desirable that the resulting certainty factor is independent of the way in which the results of the different rules are taken together to be combined. Therefore, the combination function f_{\parallel} has to respect the property of commutativity

$$f_{\parallel}(x, y) = f_{\parallel}(y, x)$$

for all certainty factors x and y , and the property of associativity

$$f_{\parallel}(f_{\parallel}(x, y), z) = f_{\parallel}(x, f_{\parallel}(y, z))$$

for all certainty factors x , y and z , as well. In addition, the combination function for co-concluding production rules should respect the property of idempotency

$$f_{\parallel}(x, x) = x$$

for all certainty factors x .

Finally, we want the four combination functions to be monotonic increasing.

Therefore, the combination functions f_{\circ} , $f_{\&}$, $f_{|}$ and $f_{||}$ have to respect the following property:

$$\begin{aligned} \text{if } x \geq x' \text{ and } y \geq y', \text{ then} \\ f_{\circ}(x, y) &\geq f_{\circ}(x', y'), \text{ and} \\ f_{\&}(x, y) &\geq f_{\&}(x', y'), \text{ and} \\ f_{|}(x, y) &\geq f_{|}(x', y'), \text{ and} \\ f_{||}(x, y) &\geq f_{||}(x', y') \end{aligned}$$

2.1.5. The Actual Combination Functions

In [SHOR84], E.H. Shortliffe and B.G. Buchanan have introduced four functions for combining certainty factors. In this section we discuss these functions and show their correspondence with our combination functions f_{\circ} , $f_{\&}$, $f_{|}$ and $f_{||}$.

The Combination Function for Propagating Uncertain Evidence.

In the case in which the evidence e in a production rule $e \rightarrow h$ is an intermediate hypothesis which has been confirmed to some degree, the certainty factor $CF(e \leftarrow u, D^{u,e})$ of the intermediate hypothesis e given some derivation of e from u is used as a weighting factor for the certainty factor $CF(h \leftarrow e, e \rightarrow h)$ associated with the hypothesis of the rule. Adapted to our notational convention, the combination function for propagating uncertain evidence proposed by Shortliffe and Buchanan reads as follows:

$$CF(h \leftarrow u, (D^{u,e}) \circ (e \rightarrow h)) = CF(h \leftarrow e, e \rightarrow h) \cdot \max\{0, CF(e \leftarrow u, D^{u,e})\}$$

or using the function f_{\circ} ,

$$f_{\circ}(x, y) = y \cdot \max\{0, x\}$$

where x denotes the certainty factor $CF(e \leftarrow u, D^{u,e})$ of the intermediate hypothesis e , and y denotes the certainty factor $CF(h \leftarrow e, e \rightarrow h)$ associated with the hypothesis h of the production rule $e \rightarrow h$. In general, this combination function does not respect the property of monotony; however, in case only non-negative certainty factors have been associated with the hypotheses of the rules, the function is monotonic increasing.

Shortliffe and Buchanan proposed the following formulation for the combination function for propagating uncertain evidence (although the function is not explicitly stated in their original work, it is the straightforward analogy of the corresponding functions for their basic measures of uncertainty which will be discussed in Section 2.2):

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

where $CF'(h, e)$ is the certainty factor associated with the hypothesis h given that the evidence e is observed with absolute certainty, that is, the certainty factor the expert has associated with the hypothesis h in the production rule $e \rightarrow h$. The certainty factor $CF(e, e')$ denotes the actual certainty factor of e given some prior evidence e' ; similarly $CF(h, e)$ is the actual certainty factor of h after the application of the rule $e \rightarrow h$. In our opinion, the actual certainty factor of h after the application of the production rule $e \rightarrow h$, expressed in the left-hand side of the formulation given above, is not only dependent upon h and e , but upon the prior evidence e' as well. This dependency on e' is not expressed in the original formulation of the combination function. It is this inadequacy that caused the necessity of introducing the seemingly strange quoted function CF' . Our observation that the actual certainty factor of the hypothesis h is dependent upon all intermediate hypotheses that were used in deriving h has led to the introduction of the notion of a derivation with respect to a fixed set of production rules into our formulation of the certainty factor function. Note that we have not affected the intended meaning of the original formulation of the combination function for propagating uncertain evidence.

The Combination Functions for Composite Hypotheses.

If the evidence e in the production rule $e \rightarrow h$ is a conjunction of disjunctions of pieces of evidence, the certainty factors of each of the constituting pieces of evidence are combined into a single certainty factor for e . Recall that for this purpose we have introduced the combination functions $f_{\&}$ and $f_{|}$. The combination function proposed by Shortliffe and Buchanan for a conjunction of hypotheses reads as follows:

$$\begin{aligned} CF(e_1 \wedge e_2 \rightarrow u, (D^{u, e_1} \& D^{u, e_2})) = \\ = \min\{CF(e_1 \rightarrow u, D^{u, e_1}), CF(e_2 \rightarrow u, D^{u, e_2})\} \end{aligned}$$

For a combination function for a disjunction of hypotheses they have chosen:

$$\begin{aligned} CF(e_1 \vee e_2 \rightarrow u, (D^{u, e_1} | D^{u, e_2})) = \\ = \max\{CF(e_1 \rightarrow u, D^{u, e_1}), CF(e_2 \rightarrow u, D^{u, e_2})\} \end{aligned}$$

Using the functions $f_{\&}$ and $f_{|}$ we have:

$$\begin{aligned} f_{\&}(x, y) &= \min\{x, y\}, \text{ and} \\ f_{|}(x, y) &= \max\{x, y\} \end{aligned}$$

where x denotes the certainty factor $CF(e_1 \rightarrow u, D^{u, e_1})$ and y denotes the certainty factor $CF(e_2 \rightarrow u, D^{u, e_2})$. From this formulation it should be evident that these combination functions respect the properties of commutativity, associativity, idempotency and monotony.

Shortliffe and Buchanan proposed the following formulation for these combination functions:

$$CF(h_1 \wedge h_2, e) = \min\{CF(h_1, e), CF(h_2, e)\}$$

$$CF(h_1 \vee h_2, e) = \max\{CF(h_1, e), CF(h_2, e)\}$$

These combination functions may be used only to combine the certainty factors of several hypotheses given the *same* evidence. In practice, however, the certainty factors of the hypotheses to be combined are generally derived along different inference paths, and differ in their second argument due to the original formulation of the combination function for propagating uncertain evidence. The reader can verify that our reformulation of the combination functions for composite hypotheses has the same meaning as the original formulation.

The Combination Function for Co-Concluding Production Rules.

The combination function which remains to be discussed concerns several different production rules concluding on the same hypothesis, that is, our function $f_{||}$. The following combination function is given to deal with this situation:

$$(1) \quad CF_{||}(h \leftarrow u, D_1^{u,h} \parallel D_2^{u,h}) = CF(h \leftarrow u, D_1^{u,h}) + CF(h \leftarrow u, D_2^{u,h}) \cdot (1 - CF(h \leftarrow u, D_1^{u,h})), \text{ if } CF(h \leftarrow u, D_1^{u,h}) > 0 \text{ and } CF(h \leftarrow u, D_2^{u,h}) > 0, \text{ and}$$

$$(2) \quad CF_{||}(h \leftarrow u, D_1^{u,h} \parallel D_2^{u,h}) = \frac{CF(h \leftarrow u, D_1^{u,h}) + CF(h \leftarrow u, D_2^{u,h})}{1 - \min\{|CF(h \leftarrow u, D_1^{u,h})|, |CF(h \leftarrow u, D_2^{u,h})|\}},$$

if $-1 < CF(h \leftarrow u, D_1^{u,h}) \cdot CF(h \leftarrow u, D_2^{u,h}) \leq 0$, and

$$(3) \quad CF_{||}(h \leftarrow u, D_1^{u,h} \parallel D_2^{u,h}) = CF(h \leftarrow u, D_1^{u,h}) + CF(h \leftarrow u, D_2^{u,h}) \cdot (1 + CF(h \leftarrow u, D_1^{u,h})), \text{ if } CF(h \leftarrow u, D_1^{u,h}) < 0 \text{ and } CF(h \leftarrow u, D_2^{u,h}) < 0.$$

Note that $CF_{||}(h \leftarrow u, D_1^{u,h} \parallel D_2^{u,h})$ is not defined for the case where $CF(h \leftarrow u, D_1^{u,h}) \cdot CF(h \leftarrow u, D_2^{u,h}) = -1$. In fact, it is not even possible to find a continuous extension of the function $CF_{||}$ for this case. The fill-in for the combination function $f_{||}$ employed in the certainty factor model therefore is not a total function.

Using our function $f_{||}$ we obtain the following more perspicuous

formulation of the actual combination function for co-concluding production rules:

$$f_{||}(x, y) = \begin{cases} x + y - xy & \text{if } x, y > 0 \\ \frac{x + y}{1 - \min\{|x|, |y|\}} & \text{if } -1 < x \cdot y \leq 0 \\ x + y + xy & \text{if } x, y < 0 \end{cases}$$

where x denotes the certainty factor $CF(h \leftarrow u, D_1^{u,h})$ and y denotes the certainty factor $CF(h \leftarrow u, D_2^{u,h})$. This combination function respects the properties of commutativity and associativity, as shown in [SPIE86a]. Furthermore, the function is monotonic increasing. However, it does not respect the property of idempotency.

In [SHOR84], the following formulation for this combination function is given:

- (1) $CF(h, e_1 \wedge e_2) = CF(h, e_1) + CF(h, e_2)(1 - CF(h, e_1))$,
if $CF(h, e_1) > 0$ and $CF(h, e_2) > 0$, and
- (2) $CF(h, e_1 \wedge e_2) = \frac{CF(h, e_1) + CF(h, e_2)}{1 - \min\{|CF(h, e_1)|, |CF(h, e_2)|\}}$,
if one of $CF(h, e_i) < 0$, $i = 1, 2$, and
- (3) $CF(h, e_1 \wedge e_2) = CF(h, e_1) + CF(h, e_2)(1 + CF(h, e_1))$,
if $CF(h, e_1) < 0$ and $CF(h, e_2) < 0$.

It is noted that this function (mistakenly) is not defined if at least one of $CF(h, e_1)$ and $CF(h, e_2)$ equals 0. Furthermore, the case in which $CF(h, e_1) \cdot CF(h, e_2) = -1$ should be excluded explicitly since the combination function is undefined in this case. A more serious criticism is that in this formulation of combining the results of production rules concluding on the same hypothesis, the same symbol \wedge is used as in describing a conjunction of two hypotheses or pieces of evidence. Shortliffe and Buchanan seem to assume that the success of a production rule $e_1 \wedge e_2 \rightarrow h$ is equivalent to the success of the two production rules $e_1 \rightarrow h$ and $e_2 \rightarrow h$. As such an equivalence is apt to be violated due to inconsistent function values given by the expert (and the user), we have introduced another notational convention. Again, our reformulation does not change the original meaning of the combination function.

To conclude with, we demonstrate the application of the combination functions by means of a numerical example.

EXAMPLE 2.10. Consider the following three production rules:

$$d \wedge f \rightarrow b$$

$$a \rightarrow h$$

$$b \wedge c \rightarrow h$$

The expert has provided the following certainty factors:

$$CF(b \dashv d \wedge f, d \wedge f \rightarrow b) = 0.80$$

$$CF(h \dashv a, a \rightarrow h) = 0.70$$

$$CF(h \dashv b \wedge c, b \wedge c \rightarrow h) = 0.50$$

We assume that h is the goal hypothesis. The user of the system supplies during the consultation the following information:

$$CF(a \dashv u, u \rightarrow a) = 0.50$$

$$CF(c \dashv u, u \rightarrow c) = 0.40$$

$$CF(d \dashv u, u \rightarrow d) = 1.00$$

$$CF(f \dashv u, u \rightarrow f) = 0.90$$

Then, it takes the following computations to arrive at a certainty factor for h :

- (1) $CF(h \dashv u, (u \rightarrow a) \circ (a \rightarrow h)) = 0.70 \cdot 0.50 = 0.35$
- (2) $CF(d \wedge f \dashv u, (u \rightarrow d) \& (u \rightarrow f)) = \min\{1.00, 0.90\} = 0.90$
- (3) $CF(b \dashv u, ((u \rightarrow d) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b)) = 0.80 \cdot 0.90 = 0.72$
- (4) $CF(b \wedge c \dashv u, (((u \rightarrow d) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b)) \& (u \rightarrow c)) = \min\{0.40, 0.72\} = 0.40$
- (5) $CF(h \dashv u, (((u \rightarrow d) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b)) \& (u \rightarrow c)) \circ (b \wedge c \rightarrow h)) = 0.50 \cdot 0.40 = 0.20$
- (6) $CF(h \dashv u, (((u \rightarrow a) \circ (a \rightarrow h)) \parallel (((u \rightarrow d) \& (u \rightarrow f)) \circ (d \wedge f \rightarrow b)) \& (u \rightarrow c)) \circ (b \wedge c \rightarrow h))) = 0.35 + 0.20 \cdot 0.65 = 0.48$

■

2.2. THE PROBABILISTIC BASIS OF THE CERTAINTY FACTOR MODEL

In the preceding section we have presented a syntactical model for handling uncertainty in a rule-based top-down reasoning expert system. This model has been used to examine the certainty factor model as it was proposed by E.H. Shortliffe and B.G. Buchanan: we have arrived at a syntactically correct reformulation of the model. This new formulation will now be the point of departure for a discussion of some of the probabilistic issues involved in the certainty factor model.

In [SHOR84], Shortliffe and Buchanan have suggested a mathematical foundation for their model in probability theory. The certainty factor function we have introduced in the preceding section is not the basic notion of uncertainty employed in the certainty factor model: this function is defined in

terms of two basic measures of uncertainty, the *measures of belief* and *disbelief*, which in turn have been defined in terms of probability theory. In this section, we introduce this probabilistic foundation of the certainty factor model. It should be noted that although Shortliffe and Buchanan have suggested a mathematical foundation for their model, they have not provided a thorough justification for it. In the Sections 2.3 and 2.4, we present a detailed analysis of this foundation.

2.2.1. Elementary Probability Theory

This section presents a brief introduction to elementary probability theory, thus providing a point of departure for the remaining sections of this chapter. We chose [HOGG78] as a basis for our introduction although any other introductory textbook will suffice.

Many kinds of investigations may be characterized in part by conceptually repeated experimentation under essentially the same conditions. Each experiment terminates in an *outcome* which cannot be predicted with certainty prior to the performance of the experiment. The non-empty collection of all possible outcomes of an experiment is called its *sample space* and is usually denoted by Ω . In this chapter, we take a sample space Ω to be a finite set. A subset e of a sample space Ω is called an *event*. If upon the performance of the experiment the outcome is in e , it is said that event e has occurred. The event that the outcome is not in e is denoted by \bar{e} and is called the *complement* of e . So, $\bar{e} = \Omega \setminus e$. The event that occurs if and only if both the events e_1 and e_2 occur, is called the *intersection* of e_1 and e_2 , denoted by $e_1 \cap e_2$. The event occurring if e_1 occurs, e_2 occurs or both e_1 and e_2 occur, is called the *union* of e_1 and e_2 , denoted by $e_1 \cup e_2$.

DEFINITION 2.11. Let Ω denote a sample space. The sets $e_1, \dots, e_n \subseteq \Omega$, $n \geq 1$, are called *disjoint* if $e_i \cap e_j = \emptyset$, $1 \leq i, j \leq n$, $i \neq j$. The events corresponding with disjoint sets are called *mutually exclusive events*.

We now define a set function Pr on a sample space Ω such that if e is a subset of Ω , then $Pr(e)$ is a real number indicating the 'probability' that the outcome of the experiment is an element of e ; the function Pr is defined axiomatically in Definition 2.12.

DEFINITION 2.12. Let Ω denote a sample space. If a number $Pr(e)$ is associated with each subset $e \subseteq \Omega$, such that

- (1) $Pr(e) \geq 0$, and
- (2) $Pr(\Omega) = 1$, and
- (3) $Pr(e_1 \cup e_2 \cup \dots) = Pr(e_1) + Pr(e_2) + \dots$,
where e_i , $i = 1, 2, \dots$, are mutually exclusive events,

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

The following lemma states some properties of a probability function. The lemma is presented without proof; its statements, however, can easily be proven using Definition 2.12.

LEMMA 2.13. *Let Ω denote a sample space. Let Pr be a probability function on Ω . Then, the following properties hold:*

- (1) *For each $e \subseteq \Omega$, we have $Pr(e) = 1 - Pr(\bar{e})$.*
- (2) *$Pr(\emptyset) = 0$.*
- (3) *For each $e_1, e_2 \subseteq \Omega$ such that $e_1 \subseteq e_2$, we have $Pr(e_1) \leq Pr(e_2)$.*
- (4) *For each $e_1, e_2 \subseteq \Omega$ such that $e_1 \subseteq e_2$, if $Pr(e_1) = Pr(e_2)$ then for each $e_3 \subseteq \Omega$ we have that $Pr(e_1 \cap e_3) = Pr(e_2 \cap e_3)$.*

In some cases we are interested only in outcomes which are in a given nonempty subset e of a sample space Ω , for instance when several pieces of information concerning the final outcome become known in the course of the actual performance of the experiment. These pieces of information are called pieces of *evidence*. Let h be an event, called the *hypothesis*. Given that an event e occurs, that is, given that the evidence e has been observed, we are interested in the degree to which this information influences $Pr(h)$, the *prior* probability of the hypothesis h . The probability of h given e is defined in the following definition.

DEFINITION 2.14. *Let Ω denote a sample space, and let Pr be a probability function on Ω . For each $h, e \subseteq \Omega$ with $Pr(e) > 0$, the conditional probability of h given e , denoted by $Pr(h | e)$, is defined as*

$$Pr(h | e) = \frac{Pr(h \cap e)}{Pr(e)}$$

In the sequel, when writing $Pr(h | e)$ we will implicitly assume $Pr(e) > 0$ unless stated otherwise. The following lemma can easily be proven using the Definitions 2.12 and 2.14.

LEMMA 2.15. *Let Ω denote a sample space, let Pr be a probability function on Ω and let e be a subset of Ω such that $Pr(e) > 0$. The conditional probabilities given e define a probability function on Ω .*

Note that Lemma 2.15 allows us to state properties of a conditional probability function given some evidence e , analogous to the properties stated in Lemma 2.13.

Let Pr be a probability function on Ω . Furthermore, let $h, e \subseteq \Omega$. It seems natural to name the event h independent of the event e when $Pr(h | e) = Pr(h)$: the prior probability of event h is not influenced by the knowledge that event e has occurred. However, this intuitive notion of

independency is asymmetric in its arguments and is not applicable in case $Pr(e) = 0$. Therefore, a slightly modified definition is given.

DEFINITION 2.16. Let Ω be a sample space and let Pr be a probability function on Ω . The events $e_1, \dots, e_n \subseteq \Omega$, $n \geq 1$, are called (mutually) independent if

$$Pr(e_{i_1} \cap \dots \cap e_{i_k}) = Pr(e_{i_1}) \cdot \dots \cdot Pr(e_{i_k})$$

for each subset $\{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$, $1 \leq k \leq n$. The events e_1, \dots, e_n are called conditionally independent given a hypothesis $h \subseteq \Omega$ if

$$Pr(e_{i_1} \cap \dots \cap e_{i_k} | h) = Pr(e_{i_1} | h) \cdot \dots \cdot Pr(e_{i_k} | h)$$

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

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

$$Pr(h | e) = \frac{Pr(h \cap e)}{Pr(e)} = \frac{Pr(h)Pr(e)}{Pr(e)} = Pr(h)$$

is satisfied.

The following theorem is known as *Bayes' Theorem*.

THEOREM 2.17. Let Pr denote a probability function on a sample space Ω . For each $h, e \subseteq \Omega$ with $Pr(e) > 0$, $Pr(h) > 0$, we have

$$Pr(h | e) = \frac{Pr(e | h)Pr(h)}{Pr(e)}$$

Note that Bayes' Theorem may be used to reverse the 'direction' of probabilities.

In the preceding we have presented a frequentist view to probability theory. We have mentioned in Chapter 1, however, that artificial intelligence researchers generally take the subjectivist view to probability theory. Recall that a subjectivist views the probability of an event as a measure of a person's belief in the occurrence of the event, given the information that person has. Note that, according to this point of view, a probability not necessarily is a statement concerning repeated experimentation: a subjectivist is also willing to assess a probability for a unique event that cannot be considered to be an outcome of a repeatable experiment. However, subjective probabilities comply with the same set of axioms as probabilities from a frequentist viewpoint do. In the sequel, we will, like many artificial intelligence researchers before us, adhere to the subjectivist point of view and take the preceding definitions and lemmas to apply to subjective probabilities which have been assessed by a domain expert.

2.2.2. A Probabilistic Interpretation for Propositions and Derivations

In the following subsections we will discuss the probabilistic foundation of the certainty factor model as it has been suggested by Shortliffe and Buchanan. Now, recall that production rules are statements concerning atomic propositions and positive Boolean combinations of atomic propositions. Furthermore, the certainty factor function we have introduced in Section 2.1 involves propositions and derivations. Probabilities on the other hand are statements concerning sets. In order to be able to discuss the certainty factor model in a probabilistic setting, we have to construct a sample space which is expressive enough to discern between all elements of \mathcal{E} ; from now on, we simply assume that we are given a fixed sample space Ω meeting the mentioned requirement (in Chapter 3 in Proposition 3.12 we will discuss a method for obtaining a 'canonical' choice of such a sample space). Furthermore, we have to define interpretation mappings from \mathcal{E} into 2^Ω and from \mathcal{D} into 2^Ω .

We define an interpretation mapping $\iota_{\mathcal{E}}: \mathcal{E} \rightarrow 2^\Omega$ by first associating with each atomic proposition $a \in \mathcal{A}$ a specific nonempty subset $\iota_{\mathcal{E}}(a)$ of the sample space Ω . For ease of exposition in the sequel we assume that for each $a \in \mathcal{A}$ we have $\iota_{\mathcal{E}}(a) \neq \Omega$. The logical conjunction then translates into the intersection set operation; the logical disjunction translates into the union set operation. We assume that the mapping $\iota_{\mathcal{E}}$ is injective (modulo logical equivalence). It will be evident that this assumption is not a restrictive one. In Figure 2.5 the basic idea of this mapping is shown.

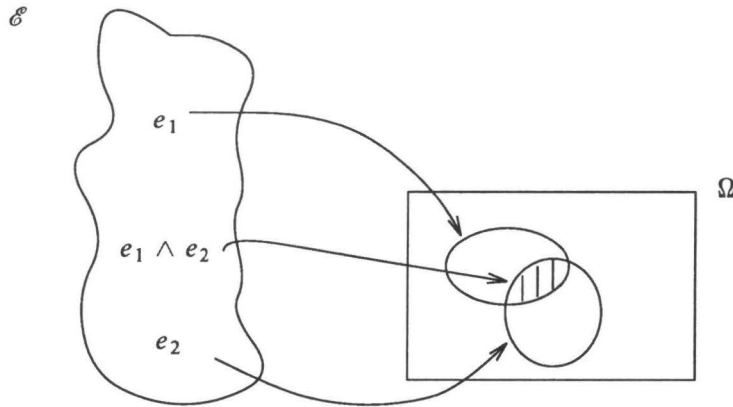


FIGURE 2.5. The interpretation mapping $\iota_{\mathcal{E}}$ from \mathcal{E} into 2^Ω .

Now, note that although we did not allow negations in production rules, each subset of Ω corresponding with a proposition nevertheless has a complement due to the properties of sets and set operations.

Similarly, each derivation $D^{i,j}$ with respect to a set of production rules \mathcal{P} is taken to identify a subset of the sample space Ω dependent upon the intermediate hypotheses that were used in deriving the hypothesis j from i . To this end, we will introduce in Definition 2.18 an interpretation mapping $\iota_{\mathcal{D}}: \mathcal{D} \rightarrow 2^{\Omega}$. The reader should bear in mind that in the foregoing we have added the notion of a derivation to the original formulation of the certainty factor model. As a consequence, in the subsequent sections our analysis of the probabilistic foundation of the model will involve the interpretation mapping $\iota_{\mathcal{D}}$. It will be evident that this mapping was not present in the probabilistic foundation as it was suggested for their model by Shortliffe and Buchanan in [SHOR84]. All results shown in the following therefore should be taken relative to the choice of this interpretation mapping. The mapping $\iota_{\mathcal{D}}$ however has been chosen after due consideration of the combination functions and the way they are applied in actual implementations of the model so as to closely fit their intended meaning.

DEFINITION 2.18. *Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Let the interpretation mapping $\iota_{\mathcal{E}}$ from \mathcal{E} into 2^{Ω} be as described above. Then, an interpretation of elements of \mathcal{D} is a mapping $\iota_{\mathcal{D}}: \mathcal{D} \rightarrow 2^{\Omega}$ such that*

- (1) *for each $u \rightarrow h \in \mathcal{P}$, we have $\iota_{\mathcal{D}}(u \rightarrow h) = \emptyset$, and*
- (2) *for each $e \rightarrow h \in \mathcal{P}$ where $e \neq u$, we have $\iota_{\mathcal{D}}(e \rightarrow h) = \iota_{\mathcal{E}}(e)$, and*
- (3) *for each $D_1, D_2 \in \mathcal{D}$, we have $\iota_{\mathcal{D}}(D_1 \circ D_2) = \iota_{\mathcal{D}}(D_1) \cup \iota_{\mathcal{D}}(D_2)$, and*
- (4) *for each $D_1, D_2 \in \mathcal{D}$, we have $\iota_{\mathcal{D}}(D_1 \& D_2) = \iota_{\mathcal{D}}(D_1) \cap \iota_{\mathcal{D}}(D_2)$, and*
- (5) *for each $D_1, D_2 \in \mathcal{D}$, we have $\iota_{\mathcal{D}}(D_1 | D_2) = \iota_{\mathcal{D}}(D_1) \cup \iota_{\mathcal{D}}(D_2)$, and*
- (6) *for each $D_1, D_2 \in \mathcal{D}$, we have $\iota_{\mathcal{D}}(D_1 \parallel D_2) = \iota_{\mathcal{D}}(D_1) \cap \iota_{\mathcal{D}}(D_2)$.*

The basic idea of this mapping $\iota_{\mathcal{D}}$ is to identify with a derivation a subset of Ω representing all information that has been concluded by the system in the course of the derivation, except for its final conclusion. So, with a derivation $u \rightarrow h$ the empty set is associated since the system has not reached any conclusions during this derivation except for h . Note that from the system's point of view a derivation $u \rightarrow h$ is a kind of 'empty' derivation in which no inference is applied. The interpretation of the conjunction and the disjunction of derivations as the set operations intersection and union respectively, is rather straightforward. The interpretation of the parallel composition of derivations as the intersection of the separate derivations, that is, the idea of taking the intersection of all evidence that is used in deriving a hypothesis, should be intuitively appealing. The proposed interpretation of the sequential composition of two derivations as the union of the sets identified by these derivations is less straightforward; it comes forth from the idea that an expert system should have the ability to extend its focus as evidence becomes available. Note that this choice renders a kind of learning behaviour of the system. In the sequel we will see that in calculating a measure of uncertainty

for a hypothesis, the set corresponding with the derivation of this hypothesis is intersected with the set identified by u . From this it will be evident that the system is not allowed to focus on hypotheses contradictory to the user's de facto knowledge.

EXAMPLE 2.19. Consider the derivation $D_1^{u,h} = ((u \rightarrow e) \circ (e \rightarrow h))$. We have $\iota_{\mathcal{D}}(D_1^{u,h}) = \iota_{\mathcal{D}}(u \rightarrow e) \cup \iota_{\mathcal{D}}(e \rightarrow h) = \emptyset \cup \iota_{\mathcal{E}}(e) = \iota_{\mathcal{E}}(e)$. Now, consider the derivation $D_2^{u,h} = (((u \rightarrow a) \circ (a \rightarrow h)) \parallel ((u \rightarrow e) \circ (e \rightarrow h)))$. We have $\iota_{\mathcal{D}}(D_2^{u,h}) = (\emptyset \cup \iota_{\mathcal{E}}(a)) \cap (\emptyset \cup \iota_{\mathcal{E}}(e)) = \iota_{\mathcal{E}}(a) \cap \iota_{\mathcal{E}}(e)$. ■

In the following, we will assume proper application of the interpretation mappings $\iota_{\mathcal{E}}$ and $\iota_{\mathcal{D}}$ implicitly as long as ambiguity cannot occur. For instance, we will write $Pr(D^{i,j})$ instead of $Pr(\iota_{\mathcal{D}}(D^{i,j}))$; for $\iota_{\mathcal{E}}(h)$ we will write \bar{h} where appropriate.

2.2.3. The Basic Measures of Uncertainty

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. These new notions of uncertainty were devised to capture the intuitive concepts of confirmation and disconfirmation, and have been inspired to a large extent by confirmation theory, [CARN50]. We will soon see that the measures of belief and disbelief have been defined in terms of probability theory; to be able to do so, Shortliffe and Buchanan have assumed the existence of a probability function Pr on their sample space. From now on, we take Pr to be a fixed probability function on Ω .

Before stating the formal definitions of the basic measures of uncertainty, we quote the intuitive account Shortliffe and Buchanan have given for the measure of belief (see also Figure 2.6).

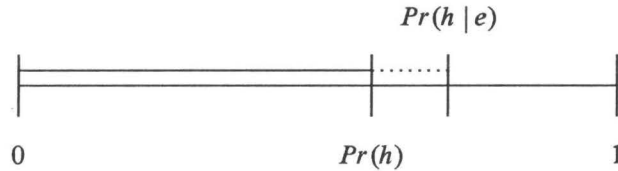


FIGURE 2.6. The degree of increased belief.

“In accordance with subjective probability theory, it may be argued that the expert's personal probability $Pr(h)$ reflects his or her belief in h at any given time. Thus, $1 - Pr(h)$ can be viewed as an estimate of the expert's disbelief

regarding the truth of h . If $Pr(h | e)$ is greater than $Pr(h)$ the observation of e increases the expert's belief in h while decreasing his or her disbelief regarding the truth of h . In fact the proportionate decrease in disbelief is given by the following ratio:

$$\frac{Pr(h | e) - Pr(h)}{1 - Pr(h)}$$

This ratio is called the measure of increased belief in h resulting from the observation of e ."

"The above definition may now be specified formally in terms of conditional and a priori probabilities:

$$MB(h, e) = \begin{cases} 1 & \text{if } Pr(h) = 1 \\ \frac{\max\{Pr(h | e), Pr(h)\} - Pr(h)}{\max\{1, 0\} - Pr(h)} & \text{otherwise} \end{cases}$$

"

([SHOR84], pp. 247, 248)

Note that the ratio mentioned in the first quotation is equal to

$$\frac{Pr(h | e) - Pr(h)}{1 - Pr(h)} = 1 - \frac{Pr(\bar{h} | e)}{Pr(\bar{h})}$$

Shortliffe and Buchanan use a similar argument to account for their measure of disbelief

$$\frac{Pr(h) - Pr(h | e)}{Pr(h)} = 1 - \frac{Pr(h | e)}{Pr(h)}$$

In the following definition, we (re)define the measures of belief and disbelief using our notational convention so as to capture the notion of derivation (recall that we use $Pr(h | e \cap D^{e,h})$ as an abbreviation for $Pr(\iota_{\mathcal{E}}(h) | \iota_{\mathcal{E}}(e) \cap \iota_{\mathcal{D}}(D^{e,h}))$). We feel that we have not affected the intended meanings of these basic measures of uncertainty.

DEFINITION 2.20. Let \mathcal{E} and \mathcal{D} be as before. Let $h, e \in \mathcal{E}$ and $D^{e,h} \in \mathcal{D}$. The measure of (increased) belief MB is a partial function $MB: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$MB(h \dashv e, D^{e,h}) = \begin{cases} 1 & \text{if } Pr(h) = 1 \\ \max\left\{0, \frac{Pr(h | e \cap D^{e,h}) - Pr(h)}{1 - Pr(h)}\right\} & \text{otherwise} \end{cases}$$

The measure of (increased) disbelief MD is a partial function $MD: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$MD(h \dashv e, D^{e,h}) = \begin{cases} 1 & \text{if } Pr(h) = 0 \\ \max\left\{0, \frac{Pr(h) - Pr(h|e \cap D^{e,h})}{Pr(h)}\right\} & \text{otherwise} \end{cases}$$

Note that in the previous definition the measures of belief and disbelief for user-supplied evidence $u \rightarrow h$ are not defined since in this case the conditional probability $Pr(h|u \cap \iota_{\mathcal{E}}(u \rightarrow h))$ is not defined. However, the definition of the measure of belief can easily be extended by taking

$$MB(h \dashv u, u \rightarrow h) = \begin{cases} 1 & \text{if } Pr(h) = 1 \\ \max\left\{0, \frac{Pr(h|u) - Pr(h)}{1 - Pr(h)}\right\} & \text{otherwise} \end{cases}$$

for the fictitious production rule $u \rightarrow h$; similarly, the definition of the measure of disbelief can be extended to provide for user-supplied evidence. In the following definition, we will implicitly assume that the definitions of the measures of belief and disbelief have been extended as indicated. However, the lemmas and propositions stated will not be proven separately for user-supplied evidence. We have chosen to do so in order to avoid obscuring the discussion by a profusion of mathematical detail. The reader may verify, however, that the results proven are not affected when user-supplied evidence is included.

Furthermore, it is noted that Shortliffe and Buchanan neither account for their choice for the measure of belief in the case $Pr(h) = 1$ nor for their choice for the measure of disbelief in the case $Pr(h) = 0$; their choices render the functions discontinuous.

According to Shortliffe and Buchanan the need for new notions of uncertainty arose from their observation that a domain expert often was unwilling to accept the logical implications of his probabilistic statements, such as: if $Pr(h|e) = x$ then $Pr(\bar{h}|e) = 1 - x$. They state that in the mentioned case an expert would claim that 'evidence e in favor of hypothesis h should not be construed as evidence against the hypothesis as well'. The reason that the logical implication concerning $Pr(\bar{h}|e)$ may seem counterintuitive is explained by J. Pearl as follows, [PEAR85]. The phrase 'evidence e in favor of hypothesis h ' is interpreted as stating an *increase* in the probability of the hypothesis from $Pr(h)$ to $Pr(h|e)$, with $Pr(h|e) > Pr(h)$: $Pr(h|e)$ is viewed relative to $Pr(h)$. On the other hand, in the argument of Shortliffe and Buchanan $Pr(\bar{h}|e)$ seems to be taken as an absolute probability irrespective of the prior $Pr(h)$. This somehow conveys the false idea that $Pr(h)$ increases by some positive factor. However if for example $Pr(h) = 0.9$ and $Pr(h|e) = 0.5$, then

no expert will construe this considerable decrease in the probability of \bar{h} as supporting the negation of h !

Anyhow, the measures of belief and disbelief explicitly capture the notion that a single piece of evidence cannot both favor and disfavor a single hypothesis. This property is stated more formally in the following lemma.

LEMMA 2.21. *Let \mathcal{E} , \mathcal{P} and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20. Let $h, e \in \mathcal{E}$ and let $D^{e,h} \in \mathcal{D}$. Then, the following properties hold:*

- (1) *If $e \rightarrow h \in \mathcal{P}$ and $Pr(h|e) = Pr(h)$ with $0 < Pr(h) < 1$, then $MB(h \dashv e, e \rightarrow h) = MD(h \dashv e, e \rightarrow h) = 0$.*
- (2) *If $MB(h \dashv e, D^{e,h}) > 0$, then $MD(h \dashv e, D^{e,h}) = 0$.*
- (3) *If $MD(h \dashv e, D^{e,h}) > 0$, then $MB(h \dashv e, D^{e,h}) = 0$.*

PROOF. We only prove the properties mentioned in the parts (1) and (2); the proof of part (3) is analogous to the one of part (2). Recall that we have chosen not to prove the lemma for the case of user-supplied evidence separately.

ad (1) We assume $Pr(h|e) = Pr(h)$ with $0 < Pr(h) < 1$. From Definition 2.20 we have

$$\begin{aligned} MB(h \dashv e, e \rightarrow h) &= \max \left\{ 0, \frac{Pr(h|e \cap \iota_{\mathcal{D}}(e \rightarrow h)) - Pr(h)}{1 - Pr(h)} \right\} = \\ &= \max \left\{ 0, \frac{Pr(h|e) - Pr(h)}{1 - Pr(h)} \right\} = 0 \end{aligned}$$

The property $MD(h \dashv e, e \rightarrow h) = 0$ follows by symmetry.

ad (2) We assume $MB(h \dashv e, D^{e,h}) > 0$. From Definition 2.20 we have

$$MB(h \dashv e, D^{e,h}) = \begin{cases} 1 & \text{if } Pr(h) = 1 \\ \max \left\{ 0, \frac{Pr(h|e \cap D^{e,h}) - Pr(h)}{1 - Pr(h)} \right\} & \text{otherwise} \end{cases}$$

We distinguish two cases: $Pr(h) = 1$ and $Pr(h) \neq 1$.

In case $Pr(h) = 1$ we have that

$$MD(h \dashv e, D^{e,h}) = \max \left\{ 0, \frac{Pr(h) - Pr(h|e \cap D^{e,h})}{Pr(h)} \right\} = 0$$

Now suppose that $Pr(h) \neq 1$. From $MB(h \dashv e, D^{e,h}) > 0$, we have

that $Pr(h | e \cap D^{e,h}) > Pr(h)$. Note that it follows that $Pr(h) > 0$. From these observations we have that

$$MD(h \dashv e, D^{e,h}) = \max \left\{ 0, \frac{Pr(h) - Pr(h | e \cap D^{e,h})}{Pr(h)} \right\} = 0$$

■

Part (1) of Lemma 2.21 shows that neither the belief nor the disbelief in a hypothesis h is increased by the observation of a piece of evidence independent of h . Parts (2) and (3) state that a single derivation of a hypothesis h cannot both confirm and disconfirm h .

2.2.4. The Combination Functions for the Basic Measures of Uncertainty

Employing the measures of belief and disbelief, and leaving the notion of certainty factors aside for the moment, an expert associates with the conclusion h of a production rule $e \rightarrow h$ a measure of belief $MB(h \dashv e, e \rightarrow h)$ and a measure of disbelief $MD(h \dashv e, e \rightarrow h)$; equally, a user associates with every piece of evidence e he feeds the system with, a measure of belief $MB(e \dashv u, u \rightarrow e)$ and a measure of disbelief $MD(e \dashv u, u \rightarrow e)$.

In terms of these measures of uncertainty, the objective of applying the certainty factor model in a rule-based top-down reasoning expert system is to calculate function values $MB(h \dashv u, D^{u,h})$ and $MD(h \dashv u, D^{u,h})$ for each goal hypothesis h . If the probability function Pr on the sample space Ω is known, then these function values $MB(h \dashv u, D^{u,h})$ and $MD(h \dashv u, D^{u,h})$ can be computed simply by using the probabilistic definitions of MB and MD . In many of the domains in which expert systems are employed, however, a probability function is rarely available. In the case where a probability function is not known, the function values $MB(h \dashv u, D^{u,h})$ and $MD(h \dashv u, D^{u,h})$ we are interested in cannot be calculated from the probabilistic foundation of the measures of uncertainty MB and MD . On the other hand, the expert and the user have supplied function values of MB and MD for only a few arguments: these functions have only been specified partially. It will be evident that the required function values $MB(h \dashv u, D^{u,h})$ and $MD(h \dashv u, D^{u,h})$ will in general not be among the specified ones.

The certainty factor model now provides *approximation functions* for calculating certain function values of MB and MD from the function values which are actually known to the system. We will see that these approximation functions fulfil the role of combination functions for MB and MD . In this subsection, we redefine the approximation functions given by Shortliffe and Buchanan for our measures of belief and disbelief: once more we add the notion of a derivation to the original formulation.

Note that it is only necessary to compute function values of MB having the form $MB(h \dashv u, D^{u,h})$, that is, with u for a second argument; a similar observation is made concerning MD . For calculating such function values, the approximation functions for MB and MD make use of the given derivation $D^{u,h}$ of the hypothesis h .

DEFINITION 2.22. Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the functions MB and MD be defined according to Definition 2.20 and let the functions MB_{\circ} , MD_{\circ} , $MB_{|}$, $MD_{|}$, $MB_{\&}$, $MD_{\&}$, $MB_{||}$ and $MD_{||}$ be as in the subsequent definitions. Let $h, e \in \mathcal{E}$ and let $D^{u,h} = D_1 \odot D_2 \in \mathcal{D}$ where $\odot \in \{\circ, \&, |, ||\}$. \overline{MB} is a partial function $\overline{MB}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

- (1) $\overline{MB}(h \dashv e, e \rightarrow h) = MB(h \dashv e, e \rightarrow h)$, if $e \rightarrow h \in \mathcal{P}$, and
- (2) $\overline{MB}(h \dashv u, D_1 \odot D_2) = MB_{\odot}(h \dashv u, D_1 \odot D_2)$, otherwise.

\overline{MD} is a partial function $\overline{MD}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

- (1) $\overline{MD}(h \dashv e, e \rightarrow h) = MD(h \dashv e, e \rightarrow h)$, if $e \rightarrow h \in \mathcal{P}$, and
- (2) $\overline{MD}(h \dashv u, D_1 \odot D_2) = MD_{\odot}(h \dashv u, D_1 \odot D_2)$, otherwise.

In Definition 2.22 we have mentioned several new functions. In the remainder of this section before each of these functions is formally defined, its intended meaning is discussed in the light of rule-based top-down reasoning expert systems.

As has been mentioned before, an expert has associated the function values $MB(h \dashv e, e \rightarrow h)$ and $MD(h \dashv e, e \rightarrow h)$ with the conclusion h of a production rule $e \rightarrow h$. Recall that these function values express the degree to which the actual occurrence of evidence e influences the belief and disbelief in the hypothesis h , respectively. When using production rules, however, an intermediate hypothesis e may be confirmed to some degree $MB(e \dashv u, D^{u,e})$ not necessarily equalling +1, and disconfirmed to some degree $MD(e \dashv u, D^{u,e})$ not always equal to 0, that is, it may be the case that the truth of e is not known with certainty. After application of the production rule $e \rightarrow h$ described above therefore, we are interested in the function values $MB(h \dashv u, D^{u,e} \circ (e \rightarrow h))$ and $MD(h \dashv u, D^{u,e} \circ (e \rightarrow h))$.

The mentioned function values now are approximated from the measures of belief and disbelief attached to the production rule and the function values $MB(e \dashv u, D^{u,e})$ and $MD(e \dashv u, D^{u,e})$ computed for the intermediate hypothesis e . In the functions for dealing with the situation that the truth of a piece of evidence is not known with certainty, the (approximated) measures of belief and disbelief of the intermediate hypothesis e are used as part of a weighting factor for the measures of belief and disbelief associated with the hypothesis h in the production rule. Note that these approximation functions act as the combination functions for uncertain evidence. These functions are denoted by \overline{MB}_{\circ} and \overline{MD}_{\circ} analogous to the notational convention introduced in the previous section; they will be called the combination functions for propagating uncertain evidence.

DEFINITION 2.23. Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the functions \overline{MB} and \overline{MD} be defined according to Definition 2.22. Let $h, e \in \mathcal{E}$, $D^{u,e} \in \mathcal{D}$ and $e \rightarrow h \in \mathcal{P}$. MB_{\circ} is a partial function $MB_{\circ}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$\begin{aligned} MB_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= \overline{MB}(h \dashv e, e \rightarrow h) \cdot \max \left\{ 0, \frac{\overline{MB}(e \dashv u, D^{u,e}) - \overline{MD}(e \dashv u, D^{u,e})}{1 - \min\{\overline{MB}(e \dashv u, D^{u,e}), \overline{MD}(e \dashv u, D^{u,e})\}} \right\} \end{aligned}$$

MD_{\circ} is a partial function $MD_{\circ}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$\begin{aligned} MD_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= \overline{MD}(h \dashv e, e \rightarrow h) \cdot \max \left\{ 0, \frac{\overline{MB}(e \dashv u, D^{u,e}) - \overline{MD}(e \dashv u, D^{u,e})}{1 - \min\{\overline{MB}(e \dashv u, D^{u,e}), \overline{MD}(e \dashv u, D^{u,e})\}} \right\} \end{aligned}$$

Note that these combination functions yield function values equal to zero when there is more reason to believe that e is false than there is to believe that e is true. This property of the combination functions for propagating uncertain evidence has as a consequence that a production rule $e \rightarrow h$ has no influence on the belief nor on the disbelief in h when the rule has failed during the top-down inference process.

Now recall that the evidence e in a production rule $e \rightarrow h$ is a positive Boolean combination of pieces of evidence. In order to be able to apply the combination functions MB_{\circ} and MD_{\circ} for approximating the measures of belief and disbelief of h after the application of this rule, the measures of belief and disbelief of e given some derivation of e from u have to be known. As these function values generally are not known they are approximated from the separate measures of belief and disbelief for each of the pieces of evidence that e comprises, then viewed as hypotheses. As an intuitive account for their approximation functions, Shortliffe and Buchanan argue

“that the measure of belief in the conjunction of two hypotheses is only as good as the belief in the hypothesis that is believed less strongly, whereas ... the measure of disbelief in such a conjunction is as strong as the disbelief in the most strongly disconfirmed.”

([SHOR84], p. 256)

Complementary observations are made for disjunctions of hypotheses. In character with these contemplations, Definition 2.24 formulates the functions $MB_{\&}$, $MD_{\&}$, $MB_{|}$ and $MD_{|}$ for approximating the measures of belief and disbelief in positive Boolean combinations of hypotheses. Note that these approximation functions fulfil the role of the combination functions for composite hypotheses. In the sequel therefore, they will be denoted as such.

DEFINITION 2.24. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions \overline{MB} and \overline{MD} be defined according to Definition 2.22. Let $e_i \in \mathcal{E}$ and $D^{u,e_i} \in \mathcal{D}$, $i = 1, 2$. MB_{\perp} is a partial function $MB_{\perp}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$MB_{\perp}(e_1 \vee e_2 \dashv u, D^{u,e_1} | D^{u,e_2}) = \max\{\overline{MB}(e_1 \dashv u, D^{u,e_1}), \overline{MB}(e_2 \dashv u, D^{u,e_2})\}$$

MD_{\perp} is a partial function $MD_{\perp}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$MD_{\perp}(e_1 \vee e_2 \dashv u, D^{u,e_1} | D^{u,e_2}) = \min\{\overline{MD}(e_1 \dashv u, D^{u,e_1}), \overline{MD}(e_2 \dashv u, D^{u,e_2})\}$$

$MB_{\&}$ is a partial function $MB_{\&}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$MB_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \min\{\overline{MB}(e_1 \dashv u, D^{u,e_1}), \overline{MB}(e_2 \dashv u, D^{u,e_2})\}$$

$MD_{\&}$ is a partial function $MD_{\&}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

$$MD_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \max\{\overline{MD}(e_1 \dashv u, D^{u,e_1}), \overline{MD}(e_2 \dashv u, D^{u,e_2})\}$$

When different successful production rules $e_i \rightarrow h$ conclude on the same hypotheses h , a measure of belief $\overline{MB}(h \dashv u, D_i^{u,h})$ and a measure of disbelief $\overline{MD}(h \dashv u, D_i^{u,h})$ are calculated from each of these rules using the approximation functions \overline{MB} and \overline{MD} . The net measure of belief and the net measure of disbelief, for example for two production rules $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$, are approximated from these partial measures of belief and disbelief. Shortliffe and Buchanan account for their combination functions for combining the results of different production rules concluding on the same hypothesis as follows:

“since an MB (or MD) represents a proportionate decrease of disbelief (or belief), the MB (or MD) of a newly acquired piece of evidence should be applied proportionately to the disbelief (or belief) still remaining.”
([SHOR84], p. 256)

The next definition formulates the approximation functions MB_{\parallel} and MD_{\parallel} for dealing with co-concluding production rules. Again, these approximation functions will be named after the role they fulfil; they will be called the combination functions for combining the results of co-concluding production rules.

DEFINITION 2.25. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions \overline{MB} and \overline{MD} be defined according to Definition 2.22. Let $h \in \mathcal{E}$ and let $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$. MB_{\parallel} is a partial function $MB_{\parallel}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

- (1) $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$, if $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$, and
- (2) $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \overline{MB}(h \dashv u, D_1^{u,h}) + \overline{MB}(h \dashv u, D_2^{u,h}) \cdot (1 - \overline{MB}(h \dashv u, D_1^{u,h}))$, otherwise.

$MD_{||}$ is a partial function $MD_{||}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [0, 1]$ such that

- (1) $MD_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$, if $MB_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$, and
- (2) $MD_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \overline{MD}(h \dashv u, D_1^{u,h}) + \overline{MD}(h \dashv u, D_2^{u,h}) \cdot (1 - \overline{MD}(h \dashv u, D_1^{u,h}))$, otherwise.

When closely examining the definitions of $MB_{||}$ and $MD_{||}$, a circularity is readily detected: the function values $MB_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ are not defined uniquely in for example the case where $\overline{MB}(h \dashv u, D_1^{u,h}) = 1$ and $\overline{MD}(h \dashv u, D_2^{u,h}) = 1$. In Definition 2.25 we have merely followed Shortliffe and Buchanan. In Section 2.3.1, however, we comment on this observation and show that, given the probabilistic foundation of the model, such problematic cases cannot occur.

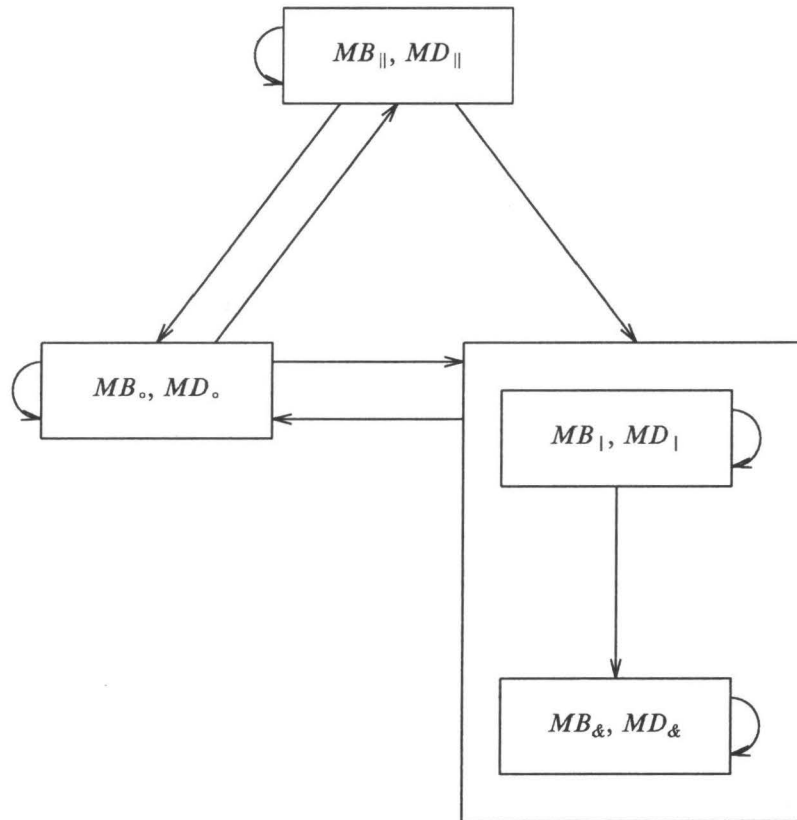


FIGURE 2.7. The order in applying the combination functions.

In the preceding, we have defined the functions \overline{MB} and \overline{MD} recursively through eight combination functions. Since in top-down inference a derivation is built up from successful production rules, these combination functions are not applied in just any order for approximating the measures of belief and disbelief for the intermediate results derived as the inference proceeds. Figure 2.7 schematically depicts the process of approximating function values as a derivation is being 'built up'; the directions of the arcs indicate the order in which the combination functions are applied. For example, from the left-hand side of a production rule being a conjunction of disjunctions of atomic propositions we have that the combination functions MB_{\downarrow} and MD_{\downarrow} cannot be applied right after $MB_{\&}$ and $MD_{\&}$ have been applied.

2.2.5. A Derived Measure and its Combination Functions

In addition to the measures of uncertainty MB and MD , in the certainty factor model a third measure, derived from the measures of belief and disbelief, is defined. This derived measure of uncertainty is the *certainty factor function* we have encountered before in Section 2.1.

DEFINITION 2.26. Let \mathcal{E} and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20. Let $h, e \in \mathcal{E}$ and $D^{e,h} \in \mathcal{D}$. The certainty factor function CF is a partial function $CF: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that

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

Recall that the measures of belief and disbelief were devised by Shortliffe and Buchanan to explicitly distinguish between the concepts of confirmation and disconfirmation. This property is 'preserved' in the certainty factor function.

LEMMA 2.27. Let \mathcal{E} and \mathcal{D} be as before. Let the functions MB and MD be defined according to Definition 2.20. Furthermore, let the function CF be as in the preceding definition. Let $h, e \in \mathcal{E}$ and $D^{e,h} \in \mathcal{D}$. Then, one of the following statements is true:

- (1) $CF(h \downarrow e, D^{e,h}) = MB(h \downarrow e, D^{e,h})$, or
- (2) $CF(h \downarrow e, D^{e,h}) = -MD(h \downarrow e, D^{e,h})$.

PROOF. From Lemma 2.21 we have that at least one of $MB(h \downarrow e, D^{e,h})$ and $MD(h \downarrow e, D^{e,h})$ equals zero. The property stated in the present lemma follows from this observation. ■

Definition 2.26 describes the certainty factor function in terms of the measures of belief and disbelief as intended by Shortliffe and Buchanan in [SHOR84]: if the exact function values $MB(h \downarrow e, D^{e,h})$ and $MD(h \downarrow e, D^{e,h})$ are known,

then the corresponding function value $CF(h \dashv e, D^{e,h})$ can be calculated from these values. As we have discussed in the preceding subsection, however, in general the function values of \overline{MB} and \overline{MD} are not known; they are approximated in the model using MB and MD . So, in practice the function values of the certainty factor function are calculated from *approximations* of the function values of MB and MD . In the following definition, therefore, the notion of a certainty factor is redefined in terms of the approximation functions for MB and MD .

DEFINITION 2.28. *Let \mathcal{E} and \mathcal{D} be as before. Furthermore, let the functions \overline{MB} and \overline{MD} be defined according to Definition 2.22. Let $h, e \in \mathcal{E}$ and $D^{e,h} \in \mathcal{D}$. CF' is a partial function $CF': \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that*

$$CF'(h \dashv e, D^{e,h}) = \frac{\overline{MB}(h \dashv e, D^{e,h}) - \overline{MD}(h \dashv e, D^{e,h})}{1 - \min\{\overline{MB}(h \dashv e, D^{e,h}), \overline{MD}(h \dashv e, D^{e,h})\}}$$

It should be evident from the definitions of the functions CF and CF' that these functions (at least) coincide where production rules are concerned. This property is formulated in the following lemma.

LEMMA 2.29. *Let \mathcal{E} and \mathcal{P} be as before. Furthermore, let the function CF be defined according to Definition 2.26 and let the function CF' be as above. Let $h, e \in \mathcal{E}$ and $e \rightarrow h \in \mathcal{P}$. Then, $CF(h \dashv e, e \rightarrow h) = CF'(h \dashv e, e \rightarrow h)$.*

PROOF. The property stated in the lemma follows from the observation that from Definition 2.22 we have $\overline{MB}(h \dashv e, e \rightarrow h) = MB(h \dashv e, e \rightarrow h)$ and $\overline{MD}(h \dashv e, e \rightarrow h) = MD(h \dashv e, e \rightarrow h)$. ■

In Section 2.4.1 we will show, however, that the two certainty factor functions do not coincide for each derivation in general.

In the implementation of the certainty factor model in the EMYCIN expert system shell derived from MYCIN and in later implementations, rather than subsequently approximating the measures of belief and disbelief for each hypothesis using \overline{MB} and \overline{MD} , and finally computing the certainty factor using Definition 2.28, only subsequently approximated certainty factors are used. This is why we never mentioned the measures of belief and disbelief in Section 2.1 where our aim was just to describe the application of the certainty factor model in a present-day rule-based setting. For the purpose of approximating the certainty factor function we introduce approximation functions for it in terms of certainty factors only.

DEFINITION 2.30. *Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the function CF' be defined according to Definition 2.28 and let the functions CF_{\circ} , $CF_{\&}$, CF_{\perp} and CF_{\parallel} be as in the*

subsequent definitions. Let $h, e \in \mathcal{E}$ and $D^{u,h} = D_1 \odot D_2 \in \mathcal{D}$ where $\odot \in \{\circ, \&, |, \parallel\}$. \overline{CF} is a partial function $\overline{CF}: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that

- (1) $\overline{CF}(h \dashv e, e \rightarrow h) = \overline{CF}'(h \dashv e, e \rightarrow h)$, if $e \rightarrow h \in \mathcal{P}$, and
- (2) $\overline{CF}(h \dashv u, D_1 \odot D_2) = \overline{CF}_\odot(h \dashv u, D_1 \odot D_2)$, otherwise.

In Definition 2.30 several new functions have been mentioned. These functions \overline{CF}_\circ , $\overline{CF}_|$, $\overline{CF}_\&$ and \overline{CF}_\parallel are the combination functions for the certainty factor function which have been discussed before informally in Section 2.1.5. They are defined more formally in the following three definitions.

DEFINITION 2.31. Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the function \overline{CF} be defined according to Definition 2.30. Let $h, e \in \mathcal{E}$, $D^{u,e} \in \mathcal{D}$ and $e \rightarrow h \in \mathcal{P}$. \overline{CF}_\circ is a partial function $\overline{CF}_\circ: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that

$$\overline{CF}_\circ(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = \overline{CF}(h \dashv e, e \rightarrow h) \cdot \max\{0, \overline{CF}(e \dashv u, D^{u,e})\}$$

Henceforth, \overline{CF}_\circ will be called the combination function for propagating uncertain evidence, analogous to the naming of MB_\circ and MD_\circ . This combination function shows once more that a production rule has no influence on the belief nor on the disbelief in a hypothesis when the rule has failed during the top-down inference process: in that case the approximated certainty factor resulting from application of this rule equals zero.

DEFINITION 2.32. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function \overline{CF} be defined according to Definition 2.30. Let $e_i \in \mathcal{E}$ and $D^{u,e_i} \in \mathcal{D}$, $i = 1, 2$. $\overline{CF}_|$ is a partial function $\overline{CF}_|: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that

$$\overline{CF}_|(e_1 \vee e_2 \dashv u, D^{u,e_1} | D^{u,e_2}) = \max\{\overline{CF}(e_1 \dashv u, D^{u,e_1}), \overline{CF}(e_2 \dashv u, D^{u,e_2})\}$$

$\overline{CF}_\&$ is a partial function $\overline{CF}_\&: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that

$$\overline{CF}_\&(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \min\{\overline{CF}(e_1 \dashv u, D^{u,e_1}), \overline{CF}(e_2 \dashv u, D^{u,e_2})\}$$

From now on, we will call the approximation functions $\overline{CF}_|$ and $\overline{CF}_\&$ the combination functions for composite hypotheses.

DEFINITION 2.33. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function \overline{CF} be defined according to Definition 2.30. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$. \overline{CF}_\parallel is a partial function $\overline{CF}_\parallel: \mathcal{E} \times \mathcal{E} \times \mathcal{D} \rightarrow [-1, 1]$ such that

- (1) $\overline{CF}_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \overline{CF}(h \dashv u, D_1^{u,h}) + \overline{CF}(h \dashv u, D_2^{u,h}) \cdot (1 - \overline{CF}(h \dashv u, D_1^{u,h}))$, if $\overline{CF}(h \dashv u, D_1^{u,h}) > 0$ and $\overline{CF}(h \dashv u, D_2^{u,h}) > 0$,

$$(2) \quad CF_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \frac{\overline{CF}(h \dashv u, D_1^{u,h}) + \overline{CF}(h \dashv u, D_2^{u,h})}{1 - \min\{|\overline{CF}(h \dashv u, D_1^{u,h})|, |\overline{CF}(h \dashv u, D_2^{u,h})|\}},$$

if $-1 < \overline{CF}(h \dashv u, D_1^{u,h}) \cdot \overline{CF}(h \dashv u, D_2^{u,h}) \leq 0$, and

$$(3) \quad CF_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \overline{CF}(h \dashv u, D_1^{u,h}) + \overline{CF}(h \dashv u, D_2^{u,h}) \cdot (1 + \overline{CF}(h \dashv u, D_1^{u,h})),$$

if $\overline{CF}(h \dashv u, D_1^{u,h}) < 0$ and $\overline{CF}(h \dashv u, D_2^{u,h}) < 0$.

The approximation function CF_{\parallel} will be called the combination function for (combining the results of) co-concluding production rules.

2.2.6. Summary of the Definitions

In the preceding subsections we have defined the basic measures of uncertainty of the certainty factor model, the measure of belief MB and the measure of disbelief MD , in terms of a probability function Pr on a sample space Ω (see Definition 2.20). As such a probability function is not always known in practice, not all function values of MB and MD can be computed from this probabilistic definition; the functions \overline{MB} and \overline{MD} are introduced to approximate function values of MB and MD respectively (see Definition 2.22). In addition, in the model a third measure of uncertainty is used; the certainty factor function CF is defined in terms of MB and MD (see Definition 2.26).

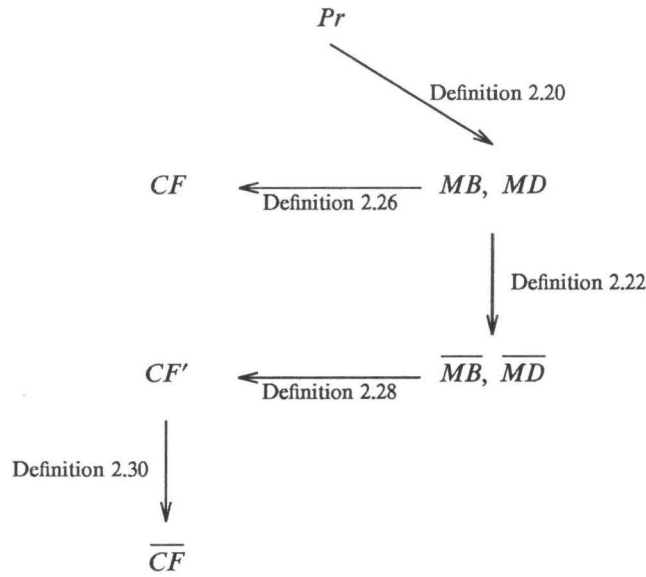


FIGURE 2.8. A diagram of the functions of the certainty factor model.

As function values of MB and MD are approximated using \overline{MB} and \overline{MD} , we have redefined the certainty factor function in terms of these approximated measures of belief and disbelief, giving CF' (see Definition 2.28). In more recent implementations of the model, only this certainty factor function is used for handling uncertainty. For the purpose of approximating function values of CF' the function \overline{CF} is introduced (see Definition 2.30). Figure 2.8 shows the relationships that have been defined between the functions employed in the certainty factor model. In the subsequent sections we examine these relationships in detail. In Section 2.3 we will concentrate on the right half of Figure 2.8 and we will show that the approximation functions \overline{MB} and \overline{MD} do not respect the probabilistic definitions of MB and MD respectively; in Section 2.4 we will investigate the left half of the figure and we will show, among other results, that the two certainty factor functions CF' and \overline{CF} coincide.

2.3. AN ANALYSIS OF THE APPROXIMATION FUNCTIONS \overline{MB} AND \overline{MD}

Since its introduction in the 1970s the certainty factor model has been implemented in a large number of rule-based expert systems and expert system shells. Part of the success of the model can be accounted for by its computational simplicity. At the same time, however, the model has been criticized severely because of its ad hoc character. In this and the following section we will show that the model does not respect the probabilistic foundation suggested for it by E.H. Shortliffe and B.G. Buchanan.

This is not the first analysis of the relationship between the certainty factor model and probability theory. J.B. Adams has examined the probabilistic basis of the model as well, [ADAM84]. In their paper [WISE86], B.P. Wise and M. Henrion suggest some properties that are implicitly assumed in the model. We will comment on these papers. D. Heckerman in [HECK86] and M. Ishizuka et al. in [ISHI81] have presented counterproposals for some parts of the model. As our only purpose is to show that the original model is not consistent with the probabilistic basis suggested for it by Shortliffe and Buchanan, we will not discuss these counterproposals.

In Section 2.2 we have defined the measures of uncertainty MB and MD in terms of probability theory. In addition, we have introduced the functions \overline{MB} and \overline{MD} for approximating certain function values of MB and MD , respectively. In this section we will analyse these approximation functions in the light of the probabilistic definitions of MB and MD . In our analysis, we will only address the question whether or not the function values of \overline{MB} and \overline{MD} obtained are exact, or more formally, whether \overline{MB} is a restriction of MB and whether \overline{MD} is a restriction of MD .

DEFINITION 2.34. Let \mathcal{U}_0 , \mathcal{U} and \mathcal{V} denote nonempty sets such that $\mathcal{U}_0 \subseteq \mathcal{U}$. Furthermore, let f be a function $f: \mathcal{U} \rightarrow \mathcal{V}$. A function $f_0: \mathcal{U}_0 \rightarrow \mathcal{V}$ is called a restriction of f , notation: $f \upharpoonright f_0$, if $f_0(u_0) = f(u_0)$ for each $u_0 \in \mathcal{U}_0$. The function f is called an extension of f_0 .

We recall that the approximation functions \overline{MB} and \overline{MD} are defined recursively through eight combination functions: MB_0 and MD_0 (the combination functions for propagating uncertain evidence), MB_1 , MD_1 , $MB_\&$ and $MD_\&$ (the combination functions for composite hypotheses), and MB_\parallel and MD_\parallel (the combination functions for co-concluding production rules). Several authors have analysed the combination functions for combining the results of co-concluding production rules, see for example [ADAM84], [HECK86], [ISHI81]. We present our views on these functions in Section 2.3.1. The other combination functions have received far less attention in the literature. We feel however that these combination functions influence the correctness of the certainty factor model as well. Practical experience in using the certainty factor model for example has learned that the functions for composite hypotheses are applied about as often as the combination functions for co-concluding production rules. The functions for composite hypotheses therefore may also have a considerable impact on the resulting approximated function values of MB and MD . In Section 2.3.3, we analyse these combination functions. Section 2.3.2 examines the combination functions for propagating uncertain evidence; it is noted that in general practice, these functions are applied less often than the other ones.

2.3.1. The Combination Functions for Co-concluding Production Rules

In this subsection, we investigate whether the combination functions for co-concluding production rules, that is, MB_\parallel and MD_\parallel , respect the probabilistic definitions of MB and MD . We are interested in function values resulting from applying the combination functions \overline{MB}_\parallel and \overline{MD}_\parallel once. Therefore, we assume that all function values of \overline{MB} and \overline{MD} which have been computed before applying MB_\parallel and MD_\parallel are exact, that is, we assume that for $i = 1, 2$ the following properties hold: $\overline{MB}(h \dashv u, D_i^{u,h}) = MB(h \dashv u, D_i^{u,h})$ and $\overline{MD}(h \dashv u, D_i^{u,h}) = MD(h \dashv u, D_i^{u,h})$.

We recall from Definition 2.25 that the combination function for combining the measures of belief of co-concluding production rules is defined as stated below:

- (1) $MB_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$, if $MD_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$, and
- (2) $MB_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MB(h, u \dashv D_1^{u,h}) + MB(h \dashv u, D_2^{u,h}) \cdot (1 - MB(h \dashv u, D_1^{u,h}))$, otherwise.

Furthermore, the combination function for combining the measures of disbelief of co-concluding production rules is defined as stated below (again assuming the above-mentioned properties):

- (1) $MD_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$, if $MB_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$, and
- (2) $MD_\parallel(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h}) \cdot (1 - MD(h \dashv u, D_1^{u,h}))$, otherwise.

We will show that in some situations under rather strong (independency) assumptions these combination functions respect the probabilistic definitions

of MB and MD . Given a hypothesis h and two different derivations $D_i^{u,h}$ of h from u each not increasing the disbelief in h , Proposition 2.35 states conditions under which the combination functions for combining these derivations are correct.

PROPOSITION 2.35. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20, and the functions MB_{\parallel} and MD_{\parallel} according to Definition 2.25. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$, such that $MD(h \dashv u, D_i^{u,h}) = 0$ and $u \cap D_i^{u,h}$ are independent and conditionally independent given h . Then,*

- (1) $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$, and
- (2) $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$.

PROOF.

ad (1) Since $MD(h \dashv u, D_1^{u,h}) = 0$ and $MD(h \dashv u, D_2^{u,h}) = 0$ imply $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) \neq 1$, we have to prove that

$$\begin{aligned} MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) \end{aligned}$$

From Definition 2.20 we have

$$\begin{aligned} MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \begin{cases} 1 & \text{if } Pr(h) = 1 \\ \max\left\{0, \frac{Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)}\right\} & \text{otherwise} \end{cases} \end{aligned}$$

We distinguish two cases: $Pr(h) = 1$ and $Pr(h) \neq 1$.

In case $Pr(h) = 1$ we have that $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MB(h \dashv u, D_1^{u,h}) = MB(h \dashv u, D_2^{u,h}) = 1$. It follows that

$$\begin{aligned} MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) = 1 \end{aligned}$$

Now suppose that $Pr(h) \neq 1$. By definition we have

$$MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \max\left\{0, \frac{Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)}\right\}$$

The fraction $\frac{Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)}$ will first be examined in isolation:

$$\begin{aligned} \frac{Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)} &= \\ &= 1 - \frac{1 - Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h})}{1 - Pr(h)} = \\ &= 1 - \frac{Pr(\bar{h} | u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(\bar{h})} = \\ &= 1 - \frac{Pr(u \cap D_1^{u,h} \cap D_2^{u,h} | \bar{h})}{Pr(u \cap D_1^{u,h} \cap D_2^{u,h})} \end{aligned}$$

using Bayes' Theorem for the last equality. We recall from the conditions of the proposition that $u \cap D_1^{u,h}$ and $u \cap D_2^{u,h}$ are independent and conditionally independent given \bar{h} . So, we have

$$\begin{aligned} \frac{Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)} &= \\ &= 1 - \frac{Pr(u \cap D_1^{u,h} | \bar{h}) Pr(u \cap D_2^{u,h} | \bar{h})}{Pr(u \cap D_1^{u,h}) Pr(u \cap D_2^{u,h})} = \\ &= 1 - \frac{Pr(\bar{h} | u \cap D_1^{u,h}) Pr(\bar{h} | u \cap D_2^{u,h})}{Pr(\bar{h})^2} \end{aligned}$$

using Bayes' Theorem once more for the last equality. The last term may now be written as follows:

$$\begin{aligned} \frac{Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)} &= \\ &= \left[1 - \frac{1 - Pr(h | u \cap D_1^{u,h})}{1 - Pr(h)} \right] + \left[1 - \frac{1 - Pr(h | u \cap D_2^{u,h})}{1 - Pr(h)} \right] + \\ &\quad - \left[1 - \frac{1 - Pr(h | u \cap D_1^{u,h})}{1 - Pr(h)} \right] \cdot \left[1 - \frac{1 - Pr(h | u \cap D_2^{u,h})}{1 - Pr(h)} \right] \end{aligned}$$

We recall from the conditions of the proposition that $MD(h \dashv u, D_1^{u,h}) = MD(h \dashv u, D_2^{u,h}) = 0$; so, for $i = 1, 2$ we have

$$\frac{Pr(h \mid u \cap D_i^{u,h}) - Pr(h)}{1 - Pr(h)} \geq 0$$

Using these inequalities it follows from Definition 2.20 that

$$\begin{aligned} \frac{Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)} &= \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) \end{aligned}$$

So, we have

$$\begin{aligned} MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \max \left\{ 0, \frac{Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)} \right\} = \\ &= \max \{ 0, MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) \} = \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) \end{aligned}$$

ad (2) We have to show that $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$ and $MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$.

If $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$ we have $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$ by definition; in case $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) \neq 1$, we have $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) = 0$ using the conditions of the proposition. It follows that $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$.

It remains to be shown that $MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$.

Since $MD(h \dashv u, D_1^{u,h}) = MD(h \dashv u, D_2^{u,h}) = 0$ implies $Pr(h) \neq 0$, we have according to Definition 2.20 that

$$MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \max \left\{ 0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} \right\}$$

We distinguish two cases: $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) > 0$ and $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$.

First, assume that $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) > 0$. The cases $Pr(h) = 1$ and $Pr(h) \neq 1$ are distinguished.

If $Pr(h) = 1$, then $Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) = 1$ as well; so,

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \max \left\{ 0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} \right\} = \\ &= 0 \end{aligned}$$

Now suppose that $Pr(h) \neq 1$. From the two assumptions $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) > 0$ and $Pr(h) \neq 1$ it follows that

$$MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \frac{Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) - Pr(h)}{1 - Pr(h)} > 0$$

From this inequality we have $Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) > Pr(h)$, implying

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \max \left\{ 0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} \right\} = \\ &= 0 \end{aligned}$$

Now assume that $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$. From this assumption we have $Pr(h) \neq 1$. Furthermore, from the proof of part (1) we have

$$\begin{aligned} MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) = 0 \end{aligned}$$

It follows that $MB(h \dashv u, D_1^{u,h}) = MB(h \dashv u, D_2^{u,h}) = 0$. This observation and the conditions of the proposition, $MD(h \dashv u, D_1^{u,h}) = 0$ and $MD(h \dashv u, D_2^{u,h}) = 0$, imply that $Pr(h \mid u \cap D_1^{u,h}) = Pr(h)$ and $Pr(h \mid u \cap D_2^{u,h}) = Pr(h)$. It can now easily be shown that $Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h}) = Pr(h)$; so,

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \max \left\{ 0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} \right\} = \\ &= 0 \end{aligned}$$

■

Given a hypothesis h and two different derivations $D_i^{u,h}$ of h from u each not increasing the belief in h , Proposition 2.36 states conditions under which the combination functions for combining these derivations respect the probabilistic definitions of MB and MD .

PROPOSITION 2.36. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20, and the functions MB_{\parallel} and MD_{\parallel} according to Definition 2.25. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$, such that $MB(h \dashv u, D_i^{u,h}) = 0$ and $u \cap D_i^{u,h}$ are independent and conditionally independent given h . Then,*

- (1) $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$, and
- (2) $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$.

PROOF. We will only prove part (1). The proof of part (2) is analogous to the proof of part (2) of the foregoing proposition.

Since $MB(h \dashv u, D_1^{u,h}) = 0$ and $MB(h \dashv u, D_2^{u,h}) = 0$ together imply that $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) \neq 1$, we have to prove that

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) \end{aligned}$$

According to Definition 2.20 we have

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \begin{cases} 1 & \text{if } Pr(h) = 0 \\ \max\left\{0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)}\right\} & \text{otherwise} \end{cases} \end{aligned}$$

We distinguish two cases: $Pr(h) = 0$ and $Pr(h) \neq 0$.

If $Pr(h) = 0$, we have by definition that $MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_1^{u,h}) = MD(h \dashv u, D_2^{u,h}) = 1$. It follows that

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) = 1 \end{aligned}$$

Now suppose that $Pr(h) \neq 0$. By definition we have

$$MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \max\left\{0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)}\right\}$$

We will examine the fraction $\frac{Pr(h) - Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)}$ in isolation.

$$\begin{aligned} \frac{Pr(h) - Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} &= \\ &= 1 - \frac{Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} = \\ &= 1 - \frac{Pr(u \cap D_1^{u,h} \cap D_2^{u,h} | h)}{Pr(u \cap D_1^{u,h} \cap D_2^{u,h})} \end{aligned}$$

using Bayes' Theorem for the last equality. Recall from the conditions of the proposition that $u \cap D_1^{u,h}$ and $u \cap D_2^{u,h}$ are independent and conditionally independent given h ; so, we have

$$\begin{aligned} \frac{Pr(h) - Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} &= \\ &= 1 - \frac{Pr(u \cap D_1^{u,h} | h) Pr(u \cap D_2^{u,h} | h)}{Pr(u \cap D_1^{u,h}) Pr(u \cap D_2^{u,h})} = \\ &= 1 - \frac{Pr(h | u \cap D_1^{u,h}) Pr(h | u \cap D_2^{u,h})}{Pr(h)^2} \end{aligned}$$

using Bayes' Theorem once more for the last equality. The last term may now be written as follows

$$\begin{aligned} \frac{Pr(h) - Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} &= \\ &= \left[1 - \frac{Pr(h | u \cap D_1^{u,h})}{Pr(h)} \right] + \left[1 - \frac{Pr(h | u \cap D_2^{u,h})}{Pr(h)} \right] + \\ &\quad - \left[1 - \frac{Pr(h | u \cap D_1^{u,h})}{Pr(h)} \right] \cdot \left[1 - \frac{Pr(h | u \cap D_2^{u,h})}{Pr(h)} \right] \end{aligned}$$

We recall from the conditions of the proposition that $MB(h \dashv u, D_1^{u,h}) = MB(h \dashv u, D_2^{u,h}) = 0$. It follows that we have

$$\frac{Pr(h) - Pr(h | u \cap D_i^{u,h})}{Pr(h)} \geq 0$$

for $i = 1, 2$. Using these inequalities and Definition 2.20 we have

$$\begin{aligned} \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} &= \\ &= MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) \end{aligned}$$

So, it follows that

$$\begin{aligned} MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \max \left\{ 0, \frac{Pr(h) - Pr(h \mid u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(h)} \right\} = \\ &= \max \{ 0, MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) \} = \\ &= MD(h \dashv u, D_1^{u,h}) + MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) \end{aligned}$$

■

Given two derivations $D_1^{u,h}$ and $D_2^{u,h}$ of h from u with respect to a given set of production rules, there are three possibilities for their relationship with the belief in the hypothesis h :

- (1) both $D_1^{u,h}$ and $D_2^{u,h}$ do not increase the disbelief in h , that is, $MD(h \dashv u, D_1^{u,h}) = MD(h \dashv u, D_2^{u,h}) = 0$, and $MB(h \dashv u, D_1^{u,h}) \geq 0$ and $MB(h \dashv u, D_2^{u,h}) \geq 0$, or
- (2) both $D_1^{u,h}$ and $D_2^{u,h}$ do not increase the belief in h , that is, $MB(h \dashv u, D_1^{u,h}) = MB(h \dashv u, D_2^{u,h}) = 0$, and $MD(h \dashv u, D_1^{u,h}) \geq 0$ and $MD(h \dashv u, D_2^{u,h}) \geq 0$, or
- (3) one of $D_1^{u,h}$ and $D_2^{u,h}$ increases the disbelief in h while the other one increases the belief in h , that is, we have that $MB(h \dashv u, D_1^{u,h}) > 0$ and $MD(h \dashv u, D_2^{u,h}) > 0$, or alternatively that $MD(h \dashv u, D_1^{u,h}) > 0$ and $MB(h \dashv u, D_2^{u,h}) > 0$.

In Proposition 2.35 it has been shown that in the case of part (1) the (approximated) function values $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ equal the actual function values of MB and MD if certain conditions are fulfilled; similarly Proposition 2.36 provides for the case that part (2) occurs. The case of ‘conflicting’ derivations described in part (3) has not been dealt with as yet.

In his analysis of the combination functions for co-concluding production rules [ADAM84], Adams states propositions similar to our Propositions 2.35 and 2.36. He, however, does not identify the restrictions $MD(h \dashv u, D_i^{u,h}) = 0$

on the values of the measure of disbelief as being necessary for showing that the combination function MB_{\parallel} respects the probabilistic definition of MB in the case of part (1); equally, he does not identify the restriction $MB(h \dashv u, D_1^{u,h}) = 0$ as being necessary for showing that MD_{\parallel} respects the probabilistic definition of MD in the case of part (2). In considering the case that part (3) occurs, he then seems to assume the following three properties:

- (i) $D_1^{u,h}$ and $D_2^{u,h}$ are mutually independent, and
- (ii) $D_1^{u,h}$ and $D_2^{u,h}$ are conditionally independent given h , and
- (iii) $D_1^{u,h}$ and $D_2^{u,h}$ are conditionally independent given \bar{h} .

It can easily be shown that when these properties hold, at least one of the following statements is true:

- (i) $Pr(h) = 0$, or
- (ii) $Pr(h) = 1$, or
- (iii) $Pr(h | D_1^{u,h}) = Pr(h)$, or
- (iv) $Pr(h | D_2^{u,h}) = Pr(h)$.

Therefore, his taking the three assumptions together renders the combination functions MB_{\parallel} and MD_{\parallel} only correct in trivial situations.

The following example concerns the case of conflicting derivations.

EXAMPLE 2.37. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20, and let MB_{\parallel} and MD_{\parallel} be as above. Let $h \in \mathcal{E}$ and $D_1^{u,h}, D_2^{u,h} \in \mathcal{D}$ such that $MB(h \dashv u, D_1^{u,h}) > 0$ and $MD(h \dashv u, D_2^{u,h}) > 0$. Now consider the exact function values $MB(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ of MB and MD respectively: from Lemma 2.21 we have that one of them equals zero. Application of the approximation functions MB_{\parallel} and MD_{\parallel} however, may render function values $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ both being greater than zero. For example, if $MB(h \dashv u, D_1^{u,h}) = 0.3$ and $MD(h \dashv u, D_2^{u,h}) = 0.4$ (and therefore $MB(h \dashv u, D_2^{u,h}) = 0$ and $MD(h \dashv u, D_1^{u,h}) = 0$), we find $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0.3$ and $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0.4$. Note that, as the approximation functions cannot decrease the once calculated measures of belief and disbelief, such an error cannot be reduced; only if one of $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ attains the value one, is the other set to zero. ■

From the preceding example, it will be evident that the approximation functions MB_{\parallel} and MD_{\parallel} do not respect the probabilistic definitions of MB and MD in the case that there are conflicting derivations of a hypothesis.

We conclude this subsection with two more propositions concerning the cases that have been treated as exceptional ones by Shortliffe and Buchanan. In [SHOR84], the case in which $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$ has been defined as an exceptional case for the function MB_{\parallel} ; likewise, the case in which $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 1$ has been defined as an exceptional case for the

function $MD_{||}$. In Definition 2.25 we have followed Shortliffe and Buchanan in excepting these cases. Recall that the function values $MB_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ and $MD_{||}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$ are not defined uniquely in for example the case where $MB(h \dashv u, D_1^{u,h}) = 1$ and $MD(h \dashv u, D_2^{u,h}) = 1$, that is, they are not defined uniquely in the case where one derivation completely proves a hypothesis h and the other one completely disconfirms h . Propositions 2.38 and 2.39 however show that under the conditions stated in the preceding propositions, the special cases mentioned above cannot occur.

PROPOSITION 2.38. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$, such that $u \cap D_i^{u,h}$ are independent and conditionally independent given h . If $MB(h \dashv u, D_1^{u,h}) = 1$, then $MD(h \dashv u, D_1^{u,h}) = MD(h \dashv u, D_2^{u,h}) = 0$.*

PROOF. From $MB(h \dashv u, D_1^{u,h}) = 1$ and Lemma 2.21 it follows that $MD(h \dashv u, D_1^{u,h}) = 0$. It is noted that from $MD(h \dashv u, D_1^{u,h}) = 0$ we have that $Pr(h) \neq 0$. We now have to show that from $MB(h \dashv u, D_1^{u,h}) = 1$ it follows that $MD(h \dashv u, D_2^{u,h}) = 0$ for any other derivation $D_2^{u,h}$ of h .

We distinguish two cases: $Pr(h) = 1$ and $Pr(h) \neq 1$.

If $Pr(h) = 1$, then also $MB(h \dashv u, D_2^{u,h}) = 1$ according to Definition 2.20. Using Lemma 2.21 once more we have $MD(h \dashv u, D_2^{u,h}) = 0$.

Now suppose that $Pr(h) \neq 1$. According to Definition 2.20 we have

$$MB(h \dashv u, D_1^{u,h}) = \max \left\{ 0, \frac{Pr(h | u \cap D_1^{u,h}) - Pr(h)}{1 - Pr(h)} \right\}$$

From $MB(h \dashv u, D_1^{u,h}) = 1$ and our assumption $Pr(h) \neq 1$ it follows that $Pr(h | u \cap D_1^{u,h}) = 1$.

From $Pr(h | u \cap D_1^{u,h}) = \frac{Pr(h \cap u \cap D_1^{u,h})}{Pr(u \cap D_1^{u,h})} = 1$ and Lemma 2.13(4), we have

$$Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) = \frac{Pr(h \cap u \cap D_1^{u,h} \cap D_2^{u,h})}{Pr(u \cap D_1^{u,h} \cap D_2^{u,h})} = 1$$

(It is noted that we may assume that $Pr(u \cap D_1^{u,h} \cap D_2^{u,h}) \neq 0$).

It follows from Bayes' Theorem that

$$Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) = \frac{Pr(u \cap D_1^{u,h} \cap D_2^{u,h} | h) Pr(h)}{Pr(u \cap D_1^{u,h} \cap D_2^{u,h})}$$

We recall that $u \cap D_1^{u,h}$ and $u \cap D_2^{u,h}$ are independent and conditionally independent given h ; so, we have

$$Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) = \frac{Pr(u \cap D_1^{u,h} | h) Pr(u \cap D_2^{u,h} | h) Pr(h)}{Pr(u \cap D_1^{u,h}) Pr(u \cap D_2^{u,h})}$$

Using $Pr(h | u \cap D_1^{u,h}) = \frac{Pr(u \cap D_1^{u,h} | h) Pr(h)}{Pr(u \cap D_1^{u,h})} = 1$, we have

$$Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) = \frac{Pr(u \cap D_2^{u,h} | h)}{Pr(u \cap D_2^{u,h})} = \frac{Pr(h | u \cap D_2^{u,h})}{Pr(h)}$$

From $Pr(h | u \cap D_1^{u,h} \cap D_2^{u,h}) = 1$, it follows that $Pr(h | u \cap D_2^{u,h}) = Pr(h)$. So, $MB(h \dashv u, D_2^{u,h}) = 0$ and $MD(h \dashv u, D_2^{u,h}) = 0$ by definition. ■

PROPOSITION 2.39. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$, such that $u \cap D_i^{u,h}$ are independent and conditionally independent given \bar{h} . If $MD(h \dashv u, D_1^{u,h}) = 1$, then $MB(h \dashv u, D_1^{u,h}) = MB(h \dashv u, D_2^{u,h}) = 0$.*

PROOF. Analogous to the proof of Proposition 2.38. ■

2.3.2. The Combination Functions for Propagating Uncertain Evidence

In this subsection, we investigate whether the combination functions for propagating uncertain evidence, that is, the functions MB_\circ and MD_\circ , respect the probabilistic definitions of the measures of uncertainty MB and MD . We are interested in the error introduced by applying the combination functions MB_\circ and MD_\circ once. Therefore, we assume that all function values of MB and MD which are computed before applying MB_\circ and MD_\circ are exact; more in specific, we assume that for an intermediate hypothesis e used in a production rule $e \rightarrow h$ the properties $\overline{MB}(e \dashv u, D^{u,e}) = MB(e \dashv u, D^{u,e})$ and $\overline{MD}(e \dashv u, D^{u,e}) = MD(e \dashv u, D^{u,e})$ hold.

We recall from Definition 2.23 that the combination functions for propagating uncertain evidence are defined as stated below:

$$\begin{aligned} MB_\circ(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= MB(h \dashv e, e \rightarrow h) \cdot \max \left\{ 0, \frac{MB(e \dashv u, D^{u,e}) - MD(e \dashv u, D^{u,e})}{1 - \min\{MB(e \dashv u, D^{u,e}), MD(e \dashv u, D^{u,e})\}} \right\} \end{aligned}$$

and

$$\begin{aligned} MD_\circ(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= MD(h \dashv e, e \rightarrow h) \cdot \max \left\{ 0, \frac{MB(e \dashv u, D^{u,e}) - MD(e \dashv u, D^{u,e})}{1 - \min\{MB(e \dashv u, D^{u,e}), MD(e \dashv u, D^{u,e})\}} \right\} \end{aligned}$$

Under assumption of the above-mentioned properties the formulations of the functions MB_0 and MD_0 can be simplified to

$$\begin{aligned} MB_0(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) \\ MD_0(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= MD(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) \end{aligned}$$

using the property stated in Lemma 2.40 given below. Note the asymmetry in these functions.

LEMMA 2.40. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20. Let $e \in \mathcal{E}$ and $D^{u,e} \in \mathcal{D}$. Then,*

$$\max \left\{ 0, \frac{MB(e \dashv u, D^{u,e}) - MD(e \dashv u, D^{u,e})}{1 - \min\{MB(e \dashv u, D^{u,e}), MD(e \dashv u, D^{u,e})\}} \right\} = MB(e \dashv u, D^{u,e})$$

PROOF. From Lemma 2.21 we have that at least one of $MB(e \dashv u, D^{u,e})$ and $MD(e \dashv u, D^{u,e})$ equals zero. The property stated in the present lemma follows from this observation. ■

From Example 2.37 it will be evident that this simplifying property does not hold in general for approximated function values.

In [ADAM84], Adams notices the resemblance between the function MB_0 and the probabilistic formula $Pr(h | e) = Pr(h | i) Pr(i | e)$ which holds in case $h \subseteq i \subseteq e$. He states that this assumption is not strong enough to prove that the combination functions for propagating uncertain evidence are correct with respect to the probabilistic definitions of the measures of belief and disbelief. Proposition 2.41, however, shows that Adams' observation stated above is useful. It is noted that Proposition 2.41 uses a property of the function $\iota_{\mathcal{D}}$, embedding derivations in the sample space Ω : the interpretation of the operation \circ as the set operation union is essential to the result stated in the proposition.

PROPOSITION 2.41. *Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the functions MB and MD be defined according to Definition 2.20 and the functions MB_0 and MD_0 according to Definition 2.23. Let $h, e \in \mathcal{E}$, $D^{u,e} \in \mathcal{D}$ and $e \rightarrow h \in \mathcal{P}$ such that $h \subseteq e \subseteq u \cap D^{u,e}$. Then,*

- (1) $MB_0(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = MB(h \dashv u, D^{u,e} \circ (e \rightarrow h))$, and
- (2) $MD_0(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = MD(h \dashv u, D^{u,e} \circ (e \rightarrow h))$.

PROOF. We will only prove the property stated in part (1); part (2) follows by symmetry.

From Definition 2.23 and Lemma 2.40 it follows that we have to prove that $MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e})$.

From Definition 2.20 we have

$$\begin{aligned} MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= \begin{cases} 1 & \text{if } Pr(h) = 1 \\ \max\left\{0, \frac{Pr(h | u \cap (D^{u,e} \cup e)) - Pr(h)}{1 - Pr(h)}\right\} & \text{otherwise} \end{cases} \end{aligned}$$

We distinguish two cases: $Pr(h) = 1$ and $Pr(h) \neq 1$.

If $Pr(h) = 1$, then we have $MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = 1$ and $MB(h \dashv e, e \rightarrow h) = 1$ by definition. From the condition of the proposition $h \subseteq e$ and our assumption $Pr(h) = 1$, it furthermore follows that $Pr(e) = 1$, which implies $MB(e \dashv u, D^{u,e}) = 1$. So, we have that $MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) = 1$.

Now suppose that $Pr(h) \neq 1$. We have

$$\begin{aligned} MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= \max\left\{0, \frac{Pr(h | u \cap (D^{u,e} \cup e)) - Pr(h)}{1 - Pr(h)}\right\} = \\ &= \max\left\{0, \frac{Pr(h | (u \cap D^{u,e}) \cup (u \cap e)) - Pr(h)}{1 - Pr(h)}\right\} = \\ &= \max\left\{0, \frac{Pr(h | u \cap D^{u,e}) - Pr(h)}{1 - Pr(h)}\right\} \end{aligned}$$

using $e \subseteq u \cap D^{u,e}$ for the last equality. Now consider the product $MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e})$. Recall that we have to show that this product equals $MB(h \dashv u, D^{u,e} \circ (e \rightarrow h))$. Once more, we distinguish two cases: $Pr(e) = 1$ and $Pr(e) \neq 1$.

If $Pr(e) = 1$, then we have $MB(e \dashv u, D^{u,e}) = 1$ by definition. Furthermore, we have

$$MB(h \dashv e, e \rightarrow h) = \max\left\{0, \frac{Pr(h | e) - Pr(h)}{1 - Pr(h)}\right\} = 0$$

From the condition of the proposition $e \subseteq u \cap D^{u,e}$ and our assumption $Pr(e) = 1$, it furthermore follows that $Pr(u \cap D^{u,e}) = 1$. We therefore have

$$MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = \max \left\{ 0, \frac{Pr(h | u \cap D^{u,e}) - Pr(h)}{1 - Pr(h)} \right\} = 0$$

So, $MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) = MB(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = 0$.

Now suppose that $Pr(e) \neq 1$. We have

$$\begin{aligned} MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) &= \\ &= \max \left\{ 0, \frac{Pr(h | e) - Pr(h)}{1 - Pr(h)} \right\} \cdot \max \left\{ 0, \frac{Pr(e | u \cap D^{u,e}) - Pr(e)}{1 - Pr(e)} \right\} = \\ &= \max \left\{ 0, \left[\frac{Pr(h | e) - Pr(h)}{1 - Pr(h)} \right] \cdot \left[\frac{Pr(e | u \cap D^{u,e}) - Pr(e)}{1 - Pr(e)} \right] \right\} \end{aligned}$$

Note that for the last equality we have used that $e \subseteq u \cap D^{u,e}$. Now consider the product $\left[\frac{Pr(h | e) - Pr(h)}{1 - Pr(h)} \right] \cdot \left[\frac{Pr(e | u \cap D^{u,e}) - Pr(e)}{1 - Pr(e)} \right]$ first in isolation:

$$\begin{aligned} &\left[\frac{Pr(h | e) - Pr(h)}{1 - Pr(h)} \right] \cdot \left[\frac{Pr(e | u \cap D^{u,e}) - Pr(e)}{1 - Pr(e)} \right] = \\ &= \frac{Pr(h | e) Pr(e | u \cap D^{u,e}) - Pr(h | e) Pr(e) - Pr(h) Pr(e | u \cap D^{u,e}) + Pr(h) Pr(e)}{(1 - Pr(h))(1 - Pr(e))} = \\ &= \frac{Pr(h | u \cap D^{u,e}) - Pr(h) - Pr(h) Pr(e | u \cap D^{u,e}) + Pr(h) Pr(e)}{(1 - Pr(h))(1 - Pr(e))} = \\ &= \frac{(Pr(h | u \cap D^{u,e}) - Pr(h))(1 - Pr(e)) + Pr(h | u \cap D^{u,e}) Pr(e) - Pr(h) Pr(e | u \cap D^{u,e})}{(1 - Pr(h))(1 - Pr(e))} = \\ &= \frac{Pr(h | u \cap D^{u,e}) - Pr(h)}{1 - Pr(h)} + \frac{Pr(h | u \cap D^{u,e}) Pr(e) - Pr(h) Pr(e | u \cap D^{u,e})}{Pr(u \cap D^{u,e})(1 - Pr(h))(1 - Pr(e))} = \\ &= \frac{Pr(h | u \cap D^{u,e}) - Pr(h)}{1 - Pr(h)} \end{aligned}$$

So, we have

$$\begin{aligned}
 MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) &= \\
 &= \max \left\{ 0, \frac{Pr(h \mid u \cap D^{u,e}) - Pr(h)}{1 - Pr(h)} \right\} = \\
 &= MB(h \dashv u, D^{u,e} \circ (e \rightarrow h))
 \end{aligned}$$

■

2.3.3. The Combination Functions for Composite Hypotheses

In this subsection, we investigate whether the combination functions for composite hypotheses, i.e. the functions MB_{\mid} , MD_{\mid} , $MB_{\&}$ and $MD_{\&}$, respect the probabilistic definitions of MB and MD . Again we are interested in the error introduced by applying these combination functions once. We therefore assume that all function values of \overline{MB} and \overline{MD} which have been computed before applying the combination functions for composite hypotheses are exact, that is, we assume that the properties $\overline{MB}(e_i \dashv u, D^{u,e_i}) = MB(e_i \dashv u, D^{u,e_i})$ and $\overline{MD}(e_i \dashv u, D^{u,e_i}) = MD(e_i \dashv u, D^{u,e_i})$, $i = 1, 2$, hold.

We recall from Definition 2.24 that the combination functions for composite hypotheses are defined as stated below:

$$MB_{\mid}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = \max\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\}$$

and

$$MD_{\mid}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = \min\{MD(e_1 \dashv u, D^{u,e_1}), MD(e_2 \dashv u, D^{u,e_2})\}$$

and

$$MB_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \min\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\}$$

and

$$MD_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \max\{MD(e_1 \dashv u, D^{u,e_1}), MD(e_2 \dashv u, D^{u,e_2})\}$$

The combination functions for composite hypotheses have received little attention in papers dealing with the certainty factor model. Adams shows that these combination functions in general are not consistent with the probabilistic definitions of MB and MD by giving a counterexample ([ADAM84], p. 258). The idea of his counterexample concerning $MB_{\&}$ is reflected in Example 2.42.

EXAMPLE 2.42. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function MB be defined according to Definition 2.20 and the function $MB_{\&}$ according to Definition 2.24. Let $e_1, e_2 \in \mathcal{E}$ such that $e_1 \cap e_2 = \emptyset$, and let

$D^{u,e_1}, D^{u,e_2} \in \mathcal{D}$ such that $Pr(u \cap D^{u,e_1} \cap D^{u,e_2}) > 0$. Now suppose that we have $MB(e_1 \dashv u, D^{u,e_1}) = 0.2$ and $MB(e_2 \dashv u, D^{u,e_2}) = 0.4$. From Definition 2.20 we have $MB(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = 0$ since $Pr(e_1 \cap e_2 | u \cap D^{u,e_1} \cap D^{u,e_2}) = 0$. From Definition 2.24, however, we find $MB_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \min\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\} = 0.2$. ■

Similar counterexamples may be found for the combination functions $MD_{\&}$, $MB_{|}$ and $MD_{|}$.

Adams does not examine the combination functions for composite hypotheses in further detail, because to him

“the extent or importance of the use of these (combination functions) in the employment of the model is not clear, but does not seem great”

([ADAM84], p. 258).

Since these combination functions are used in the application of each production rule of which the left-hand side is not atomic, we however feel that these combination functions might have a considerable impact on the approximated measures of belief and disbelief of the goal hypotheses.

Now observe that the combination function $MB_{\&}$ bears strong resemblance to the probabilistic formula $Pr(a \cap b) = \min\{Pr(a), Pr(b)\}$ which holds when either $a \subseteq b$ or $b \subseteq a$. Because of this similarity Wise and Henrion suggest in their paper that in the combination functions for composite hypotheses *maximum correlation* of hypotheses is assumed:

“the less probable event occurs whenever the more probable event occurs”

([WISE86], p. 73).

The following example shows that even the assumption of maximum correlation of hypotheses is not strong enough to derive $MB_{\&}$ and $MD_{\&}$ from the probabilistic definitions of MB and MD respectively.

EXAMPLE 2.43. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function MB be defined according to Definition 2.20 and the function $MB_{\&}$ according to Definition 2.24. Let $e_1, e_2 \in \mathcal{E}$ such that $e_1 \subset e_2$, $e_1 \neq \emptyset$, and let $D^{u,e_1}, D^{u,e_2} \in \mathcal{D}$. From Definition 2.20 we have

$$\begin{aligned} MB(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) &= \\ &= \max \left\{ 0, \frac{Pr(e_1 \cap e_2 | u \cap D^{u,e_1} \cap D^{u,e_2}) - Pr(e_1 \cap e_2)}{1 - Pr(e_1 \cap e_2)} \right\} = \\ &= \max \left\{ 0, \frac{Pr(e_1 | u \cap D^{u,e_1} \cap D^{u,e_2}) - Pr(e_1)}{1 - Pr(e_1)} \right\} \end{aligned}$$

not equalling $\min\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\}$ in general. Even if we further assume $D^{u,e_1} \subset D^{u,e_2}$ we cannot show that $MB_{\&}$ respects the basis of MB in probability theory since we might possibly have $MB(e_2 \dashv u, D^{u,e_2}) < MB(e_1 \dashv u, D^{u,e_1})$ in spite of $e_1 \subset e_2$. ■

We have not been able to identify a set of 'natural' assumptions under which the combination functions for composite hypotheses can be shown to be correct with respect to the probabilistic definitions of the measures of belief and disbelief.

2.3.4. Summary of the Results

In this section we have addressed the question whether the approximation functions \overline{MB} and \overline{MD} for the measures of uncertainty MB and MD respect the probabilistic definitions of these functions. Since the approximation functions are defined recursively through eight combination functions, we have analysed the application of each of these combination functions in just one step in the process of approximating the actual function values of MB and MD , that is, we have renounced errors introduced earlier during the approximation process. The analysis of some of these combination functions has helped us to formulate conditions under which the function respects the probabilistic foundation of the model. Note that such conditions have only been proven to be sufficient; we have not proven them necessary.

In Section 2.3.1 our analysis of the combination functions for co-concluding production rules, that is, MB_{\parallel} and MD_{\parallel} , given two derivations $D_i^{u,h}$, $i = 1, 2$, of the hypothesis h from the user's de facto knowledge u , has shown that these combination functions respect the probabilistic basis of the model if one of the following sets of conditions holds:

- (1) Both derivations do not increase the disbelief in the hypothesis, that is, $MD(h \dashv u, D_i^{u,h}) = 0$, and the two derivations, or to be more precise $u \cap D_i^{u,h}$, are independent and conditionally independent given the hypothesis (see Proposition 2.35).
- (2) Both derivations do not increase the belief in the hypothesis, that is, $MB(h \dashv u, D_i^{u,h}) = 0$, and the two derivations are independent and conditionally independent given the complement of the hypothesis (see Proposition 2.36).

In the case of 'conflicting' derivations the combination functions for co-concluding production rules do not always respect the probabilistic definitions of the measures of belief and disbelief (see Example 2.37).

In Section 2.3.2 our analysis of the combination functions for propagating uncertain evidence, that is, of MB_{\circ} and MD_{\circ} , given a production rule $e \rightarrow h$ and a derivation $D^{u,e}$ of e from the user's de facto knowledge u , has shown that these combination functions respect the probabilistic basis of the model if $h \subseteq e \subseteq u \cap D^{u,e}$ (see Proposition 2.41). This result shows that the combination functions MB_{\circ} and MD_{\circ} are correct in case the expert system is

only able to narrow its focus and does not have the ability to turn to hypotheses slightly outside the scope of the derivation up till that moment.

In Section 2.3.3 our analysis of the combination functions for composite hypotheses, that is, of the functions MB_{\downarrow} , MD_{\downarrow} , $MB_{\&}$ and $MD_{\&}$, has not enabled us to formulate 'natural' conditions under which these functions can be shown to be correct with respect to the probabilistic basis of the model. The easy counterexamples we have given concerning these functions (see the Examples 2.42 and 2.43), however, suggest that any set of such conditions will be violated in most practical cases.

From these observations we have that the approximation function \overline{MB} is not a restriction of the function MB . A similar statement can be made concerning MD and \overline{MD} .

THEOREM 2.44. *Let the functions MB and MD be defined according to Definition 2.20, and the functions \overline{MB} and \overline{MD} according to Definition 2.22. Then, the following statements are true:*

- (1) $MB \not\preceq \overline{MB}$.
- (2) $MD \not\preceq \overline{MD}$.

In Figure 2.9 this result has been inserted into the diagram of functions, which has been introduced before in Section 2.2.

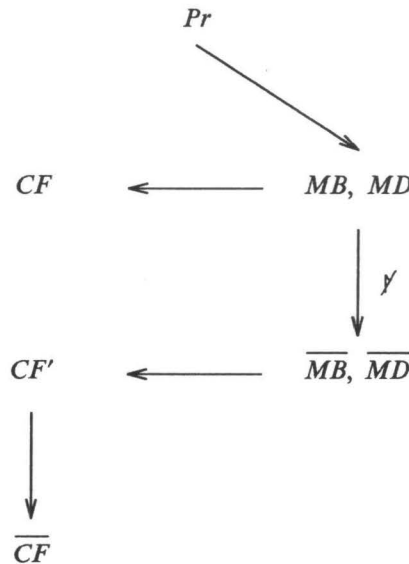


FIGURE 2.9. The diagram of functions.

It can easily be shown that the identified problems cannot be remedied, using an argument recently stated by R.E. Neapolitan, [NEAP90]. We furthermore observe that even as approximation functions for the measures of belief and disbelief, the functions MB and MD are not satisfactory; the counterexamples we have given show that these combination functions may introduce considerable errors.

2.4. AN ANALYSIS OF THE CERTAINTY FACTOR FUNCTIONS CF , CF' AND \overline{CF}

In Section 2.2.5 we have introduced in addition to the measures of belief and disbelief of the certainty factor model, a third measure of uncertainty: the certainty factor function CF . We recall from Definition 2.26 that the certainty factor function CF is defined in terms of the functions MB and MD :

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

From Lemma 2.27 we have that there is a one-to-one correspondence between the functions MB and MD , and the function CF .

Recall that, arising from the fact that in practice the function values of MB and MD are approximated using \overline{MB} and \overline{MD} , actually another certainty factor function CF' is used. Definition 2.28 redefined the certainty factor function in terms of the approximated function values of MB and MD :

$$CF'(h \vdash e, D^{e,h}) = \frac{\overline{MB}(h \vdash e, D^{e,h}) - \overline{MD}(h \vdash e, D^{e,h})}{1 - \min\{\overline{MB}(h \vdash e, D^{e,h}), \overline{MD}(h \vdash e, D^{e,h})\}}$$

Furthermore, we have described in Section 2.2.5 that in present-day implementations of the model, and in fact in all implementations since the introduction of the MYCIN system, only subsequently approximated certainty factors are used. For that purpose we have defined an approximation function \overline{CF} for certainty factors.

In this section we investigate the relationships between these certainty factor functions CF , CF' and \overline{CF} ; this section therefore focusses on the left half of Figure 2.9. From the analyses from the previous section it is readily seen that we have that application of the functions CF and CF' does not always render the same function values; in Section 2.4.1 we will discuss this observation in further detail. In Section 2.4.2 we will show that the approximation functions CF' and \overline{CF} coincide.

2.4.1. The Certainty Factor Functions CF and CF'

In this section we compare the certainty factor function CF as defined by E.H. Shortliffe and B.G. Buchanan, and the function CF' , actually employed by them in the implementation of the model. From the respective definitions of these functions we have that CF' is a restriction of CF if and only if \overline{MB} is a restriction of MB and \overline{MD} is a restriction of MD . So, using Theorem 2.44

we have that CF' is not a restriction of CF . This result will be stated more formally in Theorem 2.49. Before giving this theorem, we examine the behaviour of the functions CF and CF' in the respective cases of propagation of uncertain evidence, of composite hypotheses and of co-concluding production rules. In the following analysis we again have renounced errors that have been introduced earlier during the computation.

The first case we consider is the propagation of uncertain evidence.

COROLLARY 2.45. *Let \mathcal{E} , u and \mathcal{D} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the function CF be defined according to Definition 2.26 and the function CF' according to Definition 2.28. Let $h, e \in \mathcal{E}$, $D^{u,e} \in \mathcal{D}$ and $e \rightarrow h \in \mathcal{P}$ such that $h \subseteq e \subseteq u \cap D^{u,e}$. Then,*

$$CF'(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = CF(h \dashv u, D^{u,e} \circ (e \rightarrow h))$$

PROOF. From Proposition 2.40 it follows that under the conditions of the corollary we have $MB_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = MB(h \dashv u, D^{u,e} \circ (e \rightarrow h))$ and similarly $MD_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = MD(h \dashv u, D^{u,e} \circ (e \rightarrow h))$. The property stated in the corollary follows immediately from this observation. ■

Recall from Section 2.3.3 that we have not been able to identify a number of 'natural' conditions under which the combination functions for composite hypotheses can be shown to be correct with respect to the probabilistic definitions of MB and MD . From this observation we have that in the case of composite hypotheses the certainty factor functions CF and CF' will generally not render the same function values.

In the case of co-concluding production rules our observation concerning the two certainty factor functions is threefold. We consider the following three cases: the case of two derivations both not increasing the disbelief in a hypothesis h , the case of two derivations both not increasing the belief in h and the case of 'conflicting' derivations. Corollary 2.46 addresses the first of these cases; Corollary 2.47 concerns the second one.

COROLLARY 2.46. *Let \mathcal{E} , u and \mathcal{D} be as before. Let the function CF be defined according to Definition 2.26 and the function CF' according to Definition 2.28. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$, such that $CF(h \dashv u, D_i^{u,h}) \geq 0$ and $u \cap D_i^{u,h}$ are mutually independent and conditionally independent given h . Then,*

$$CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = CF(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$$

PROOF. The property stated in the corollary follows immediately from Proposition 2.35. ■

COROLLARY 2.47. Let \mathcal{E} , u and \mathcal{D} be as before. Let the function CF be defined according to Definition 2.26 and the function CF' according to Definition 2.28. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$, such that $CF(h \dashv u, D_i^{u,h}) \leq 0$ and $u \cap D_i^{u,h}$ are mutually independent and conditionally independent given h . Then,

$$CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = CF(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$$

PROOF. The property stated in the corollary follows immediately from Proposition 2.36. ■

The case that remains to be considered in our examination of the behaviour of the two certainty factor functions in case of co-concluding production rules is the case in which there is a derivation of h from u confirming h to some degree and a derivation of h from u disconfirming h to some degree. In [ADAM84], J.B. Adams observes that the model combines separately all derivations favouring a hypothesis and all derivations not favouring the hypothesis when calculating the corresponding certainty factor. Lemma 2.48 can easily be generalized to confirm his observation.

LEMMA 2.48. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the functions MB and MD be defined according to Definition 2.20 and the functions MB $_{\parallel}$ and MD $_{\parallel}$ according to Definition 2.25. Let the function CF' be defined according to Definition 2.28. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$. If $MB(h \dashv u, D_1^{u,h}) > 0$ and $MD(h \dashv u, D_2^{u,h}) > 0$, then

$$CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \frac{MB(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_2^{u,h})}{1 - \min\{MB(h \dashv u, D_1^{u,h}), MD(h \dashv u, D_2^{u,h})\}}$$

A similar property holds for the case where $MB(h \dashv u, D_2^{u,h}) > 0$ and $MD(h \dashv u, D_1^{u,h}) > 0$.

PROOF. Let $MB(h \dashv u, D_1^{u,h}) > 0$ and $MD(h \dashv u, D_2^{u,h}) > 0$; the proof for the case where $MB(h \dashv u, D_2^{u,h}) > 0$ and $MD(h \dashv u, D_1^{u,h}) > 0$ is analogous. From Definition 2.28 we have

$$\begin{aligned} CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \frac{MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) - MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})}{1 - \min\{MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}), MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})\}} \end{aligned}$$

It suffices to show under the conditions mentioned above, that

- (1) $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MB(h \dashv u, D_1^{u,h})$, and
- (2) $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_2^{u,h})$.

We will only prove part (1); part (2) follows by symmetry.

From the condition of the proposition $MD(h \dashv u, D_2^{u,h}) > 0$ and Lemma 2.21 it follows that $MB(h \dashv u, D_2^{u,h}) = 0$. From Definition 2.25 we have

$$\begin{aligned} MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) = \\ &= MB(h \dashv u, D_1^{u,h}) \end{aligned}$$

■

It should be evident from Example 2.37 that in the case of conflicting derivations the certainty factor functions CF and CF' do not always render the same result.

Theorem 2.49 states the conclusive result.

THEOREM 2.49. *Let the function CF be defined according to Definition 2.26 and the function CF' according to Definition 2.28. Then, $CF \not\equiv CF'$.*

2.4.2. The Combination Functions for Certainty Factors

Recall that in present-day implementations of the model instead of subsequently approximating the function values of MB and MD , and then computing the corresponding function value of CF' , only subsequently approximated certainty factors are used. For that purpose we have introduced the approximation function \overline{CF} for certainty factors. In this subsection we will show that CF' and \overline{CF} coincide.

We recall that the approximation function \overline{CF} is defined recursively through four combination functions: CF_{\circ} (the combination function for propagating uncertain evidence), CF_{\mid} and $CF_{\&}$ (the combination functions for composite hypotheses), and CF_{\parallel} (the combination function for co-concluding production rules). We will examine these combination functions separately, again renouncing errors that have been introduced earlier in the computation.

We recall from Definition 2.31 that the combination function for propagating uncertain evidence is defined as stated below:

$$CF_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = CF(h \dashv e, e \rightarrow h) \cdot \max\{0, CF(e \dashv u, D^{u,e})\}$$

Proposition 2.50 shows that the combination function CF_{\circ} respects the definition of the function CF' .

PROPOSITION 2.50. *Let \mathcal{E} , u and \mathcal{P} be as before and let \mathcal{D} be defined according to Definition 2.2. Furthermore, let the function CF' be defined according to Definition 2.28 and the function CF_{\circ} according to Definition 2.31. Let $h, e \in \mathcal{E}$, $D^{u,e} \in \mathcal{D}$ and $e \rightarrow h \in \mathcal{P}$. Then,*

$$CF_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = CF'(h \dashv u, D^{u,e} \circ (e \rightarrow h))$$

PROOF. From Definition 2.31 it follows that we have to show that

$$CF'(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = CF(h \dashv e, e \rightarrow h) \cdot \max\{0, CF(e \dashv u, D^{u,e})\}$$

From Definition 2.28 we have

$$\begin{aligned} CF'(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= \frac{MB_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)) - MD_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h))}{1 - \min\{MB_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h)), MD_{\circ}(h \dashv u, D^{u,e} \circ (e \rightarrow h))\}} \end{aligned}$$

It can easily be shown that the denominator of this fraction equals 1. It follows that

$$\begin{aligned} CF'(h \dashv u, D^{u,e} \circ (e \rightarrow h)) &= \\ &= MB(h \dashv e, e \rightarrow h) \cdot MB(e \dashv u, D^{u,e}) + \\ &\quad - MD(h \dashv e, e \rightarrow h) \cdot MD(e \dashv u, D^{u,e}) = \\ &= (MB(h \dashv e, e \rightarrow h) - MD(h \dashv e, e \rightarrow h)) \cdot MB(e \dashv u, D^{u,e}) = \\ &= (MB(h \dashv e, e \rightarrow h) - MD(h \dashv e, e \rightarrow h)) \cdot \\ &\quad \cdot \max\{0, MB(e \dashv u, D^{u,e}) - MD(e \dashv u, D^{u,e})\} \end{aligned}$$

using Lemma 2.21 for the last equality. Furthermore, we have $1 - \min\{MB(h \dashv e, e \rightarrow h), MD(h \dashv e, e \rightarrow h)\} = 1$ and similarly $1 - \min\{MB(e \dashv u, D^{u,e}), MD(e \dashv u, D^{u,e})\} = 1$. It follows that

$$\begin{aligned} MB(h \dashv e, e \rightarrow h) - MD(h \dashv e, e \rightarrow h) &= \\ &= \frac{MB(h \dashv e, e \rightarrow h) - MD(h \dashv e, e \rightarrow h)}{1 - \min\{MB(h \dashv e, e \rightarrow h), MD(h \dashv e, e \rightarrow h)\}} = \\ &= CF(h \dashv e, e \rightarrow h) \end{aligned}$$

and furthermore that

$$\begin{aligned} MB(e \dashv u, D^{u,e}) - MD(e \dashv u, D^{u,e}) &= \\ &= \frac{MB(e \dashv u, D^{u,e}) - MD(e \dashv u, D^{u,e})}{1 - \min\{MB(e \dashv u, D^{u,e}), MD(e \dashv u, D^{u,e})\}} = \\ &= CF(e \dashv u, D^{u,e}) \end{aligned}$$

So, $CF'(h \dashv u, D^{u,e} \circ (e \rightarrow h)) = CF(h \dashv e, e \rightarrow h) \cdot \max\{0, CF(e \dashv u, D^{u,e})\}$. ■

We recall from Definition 2.32 that the combination function for a disjunction of hypotheses is defined as stated below:

$$CF_{\downarrow}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = \max\{CF(e_1 \dashv u, D^{u,e_1}), CF(e_2 \dashv u, D^{u,e_2})\}$$

Proposition 2.51 shows that the combination function CF_{\downarrow} respects the definition of the function CF' .

PROPOSITION 2.51. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function CF' be defined according to Definition 2.28 and the function CF_{\downarrow} according to Definition 2.32. Let $e_i \in \mathcal{E}$ and $D^{u,e_i} \in \mathcal{D}$, $i = 1, 2$. Then,*

$$CF_{\downarrow}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = CF'(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2})$$

PROOF. From Definition 2.32 it follows that we have to show that

$$CF'(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = \max\{CF(e_1 \dashv u, D^{u,e_1}), CF(e_2 \dashv u, D^{u,e_2})\}$$

From Definition 2.28 we have that

$$\begin{aligned} CF'(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) &= \\ &= \frac{MB_{\downarrow}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) - MD_{\downarrow}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2})}{1 - \min\{MB_{\downarrow}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}), MD_{\downarrow}(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2})\}} \end{aligned}$$

Again, it can easily be shown that the denominator of the fraction equals 1. So, we have

$$\begin{aligned} CF'(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) &= \\ &= \max\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\} + \\ &\quad - \min\{MD(e_1 \dashv u, D^{u,e_1}), MD(e_2 \dashv u, D^{u,e_2})\} \end{aligned}$$

We distinguish several cases.

- (1) Assume $MB(e_1 \dashv u, D^{u,e_1}) = 0$ and $MB(e_2 \dashv u, D^{u,e_2}) = 0$. The case $MD(e_1 \dashv u, D^{u,e_1}) = MD(e_2 \dashv u, D^{u,e_2}) = 0$ follows by symmetry.

From our assumption and Lemma 2.21 we have $MD(e_1 \dashv u, D^{u,e_1}) \geq 0$ and $MD(e_2 \dashv u, D^{u,e_2}) \geq 0$.

Now suppose $MD(e_1 \dashv u, D^{u,e_1}) \leq MD(e_2 \dashv u, D^{u,e_2})$. The other case $MD(e_1 \dashv u, D^{u,e_1}) \geq MD(e_2 \dashv u, D^{u,e_2})$ follows by symmetry. Our assumptions together imply

$$\begin{aligned} & \max\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\} + \\ & - \min\{MD(e_1 \dashv u, D^{u,e_1}), MD(e_2 \dashv u, D^{u,e_2})\} = \\ & = -MD(e_1 \dashv u, D^{u,e_1}) = \\ & = MB(e_1 \dashv u, D^{u,e_1}) - MD(e_1 \dashv u, D^{u,e_1}) = \\ & = \max\{MB(e_1 \dashv u, D^{u,e_1}) - MD(e_1 \dashv u, D^{u,e_1}), \\ & \quad MB(e_2 \dashv u, D^{u,e_2}) - MD(e_2 \dashv u, D^{u,e_2})\} \end{aligned}$$

- (2) Assume $MB(e_1 \dashv u, D^{u,e_1}) > 0$ and $MD(e_2 \dashv u, D^{u,e_2}) > 0$. The case $MD(e_1 \dashv u, D^{u,e_1}) > 0$ and $MB(e_2 \dashv u, D^{u,e_2}) > 0$ follows by symmetry. From our assumption and Lemma 2.21 we have $MD(e_1 \dashv u, D^{u,e_1}) = 0$ and $MB(e_2 \dashv u, D^{u,e_2}) = 0$. So,

$$\begin{aligned} & \max\{MB(e_1 \dashv u, D^{u,e_1}), MB(e_2 \dashv u, D^{u,e_2})\} + \\ & - \min\{MD(e_1 \dashv u, D^{u,e_1}), MD(e_2 \dashv u, D^{u,e_2})\} = \\ & = MB(e_1 \dashv u, D^{u,e_1}) = \\ & = MB(e_1 \dashv u, D^{u,e_1}) - MD(e_1 \dashv u, D^{u,e_1}) = \\ & = \max\{MB(e_1 \dashv u, D^{u,e_1}) - MD(e_1 \dashv u, D^{u,e_1}), \\ & \quad MB(e_2 \dashv u, D^{u,e_2}) - MD(e_2 \dashv u, D^{u,e_2})\} \end{aligned}$$

From (1) and (2), we have

$$\begin{aligned} & CF'(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = \\ & = \max\{MB(e_1 \dashv u, D^{u,e_1}) - MD(e_1 \dashv u, D^{u,e_1}), \\ & \quad MB(e_2 \dashv u, D^{u,e_2}) - MD(e_2 \dashv u, D^{u,e_2})\} \end{aligned}$$

Using Lemma 2.21, it can easily be shown that $1 - \min\{MB(e_i \dashv u, D^{u,e_i}), MD(e_i \dashv u, D^{u,e_i})\} = 1$, $i = 1, 2$, from which we have

$$\begin{aligned} MB(e_i \dashv u, D^{u,e_i}) - MD(e_i \dashv u, D^{u,e_i}) &= \\ &= \frac{MB(e_i \dashv u, D^{u,e_i}) - MD(e_i \dashv u, D^{u,e_i})}{1 - \min\{MB(e_i \dashv u, D^{u,e_i}), MD(e_i \dashv u, D^{u,e_i})\}} = \\ &= CF(e_i \dashv u, D^{u,e_i}) \end{aligned}$$

Therefore, we have

$$CF'(e_1 \vee e_2 \dashv u, D^{u,e_1} \mid D^{u,e_2}) = \max\{CF(e_1 \dashv u, D^{u,e_1}), CF(e_2 \dashv u, D^{u,e_2})\}$$

■

We recall from Definition 2.32 that the combination function for conjunctions of hypotheses is defined as stated below:

$$CF_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = \min\{CF(e_1 \dashv u, D^{u,e_1}), CF(e_2 \dashv u, D^{u,e_2})\}$$

The proof of Proposition 2.52 is analogous to the proof of the foregoing proposition.

PROPOSITION 2.52. *Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function CF' be defined according to Definition 2.28 and the function $CF_{\&}$ according to Definition 2.32. Let $e_i \in \mathcal{E}$ and $D^{u,e_i} \in \mathcal{D}$, $i = 1, 2$. Then,*

$$CF_{\&}(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2}) = CF'(e_1 \wedge e_2 \dashv u, D^{u,e_1} \& D^{u,e_2})$$

The combination function that remains to be examined is the combination function for co-concluding production rules. We recall from Definition 2.33 that this combination function is defined as stated below:

- (1) $CF_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = CF(h \dashv u, D_1^{u,h}) + CF(h \dashv u, D_2^{u,h}) \cdot (1 - CF(h \dashv u, D_1^{u,h}))$, if $CF(h \dashv u, D_1^{u,h}) > 0$ and $CF(h \dashv u, D_2^{u,h}) > 0$,
- (2) $CF_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \frac{CF(h \dashv u, D_1^{u,h}) + CF(h \dashv u, D_2^{u,h})}{1 - \min\{|CF(h \dashv u, D_1^{u,h})|, |CF(h \dashv u, D_2^{u,h})|\}}$, if $-1 < CF_{\parallel}(h \dashv u, D_1^{u,h}) \cdot CF_{\parallel}(h \dashv u, D_2^{u,h}) \leq 0$, and

$$(3) \quad CF_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = CF(h \dashv u, D_1^{u,h}) + CF(h \dashv u, D_2^{u,h}) \cdot (1 + CF(h \dashv u, D_1^{u,h})), \text{ if } CF(h \dashv u, D_1^{u,h}) < 0 \text{ and } CF(h \dashv u, D_2^{u,h}) < 0.$$

In Proposition 2.53 it is shown that the combination function CF_{\parallel} respects the definition of the function CF' .

PROPOSITION 2.53. Let \mathcal{E} , u and \mathcal{D} be as before. Furthermore, let the function CF' be defined according to Definition 2.28 and the function CF_{\parallel} according to Definition 2.33. Let $h \in \mathcal{E}$ and $D_i^{u,h} \in \mathcal{D}$, $i = 1, 2$. Then,

$$CF_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})$$

PROOF. From Definition 2.28 we have

$$\begin{aligned} CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \frac{MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) - MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})}{1 - \min\{MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}), MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})\}} \end{aligned}$$

We will consider this fraction in detail.

- (1) Assume $MB(h \dashv u, D_1^{u,h}) > 0$ and $MB(h \dashv u, D_2^{u,h}) > 0$. The case $MD(h \dashv u, D_1^{u,h}) > 0$ and $MD(h \dashv u, D_2^{u,h}) > 0$ follows by symmetry.

From Lemma 2.21 we have $MD(h \dashv u, D_1^{u,h}) = MD(h \dashv u, D_2^{u,h}) = 0$. So, from our assumptions it follows that $CF(h \dashv u, D_1^{u,h}) > 0$ and $CF(h \dashv u, D_2^{u,h}) > 0$.

From $MD(h \dashv u, D_1^{u,h}) = 0$ and $MD(h \dashv u, D_2^{u,h}) = 0$ we have $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = 0$. Therefore, the denominator of the fraction shown above equals 1. It follows that

$$\begin{aligned} CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) - MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h}) + \\ &\quad - MB(h \dashv u, D_1^{u,h}) \cdot MB(h \dashv u, D_2^{u,h}) = \\ &= (MB(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_1^{u,h})) + (MB(h \dashv u, D_2^{u,h}) + \\ &\quad - MD(h \dashv u, D_2^{u,h})) - (MB(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_1^{u,h})) \cdot \\ &\quad \cdot (MB(h \dashv u, D_2^{u,h}) - MD(h \dashv u, D_2^{u,h})) \end{aligned}$$

From $1 - \min\{MB(h \dashv u, D_i^{u,h}), MD(h \dashv u, D_i^{u,h})\} = 1$, $i = 1, 2$, we have

$$\begin{aligned} MB(h \dashv u, D_i^{u,h}) - MD(h \dashv u, D_i^{u,h}) &= \\ &= \frac{MB(h \dashv u, D_i^{u,h}) - MD(h \dashv u, D_i^{u,h})}{1 - \min\{MB(h \dashv u, D_i^{u,h}), MD(h \dashv u, D_i^{u,h})\}} = \\ &= CF(h \dashv u, D_i^{u,h}) \end{aligned}$$

Therefore, we have

$$\begin{aligned} CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= CF(h \dashv u, D_1^{u,h}) + CF(h \dashv u, D_2^{u,h})(1 - CF(h \dashv u, D_2^{u,h})) \end{aligned}$$

- (2) Now assume $MB(h \dashv u, D_1^{u,h}) = 0$ and $MB(h \dashv u, D_2^{u,h}) > 0$. The case $MB(h \dashv u, D_1^{u,h}) > 0$ and $MB(h \dashv u, D_2^{u,h}) = 0$, similar cases for MD and the case where $MB(h \dashv u, D_1^{u,h})$, $MB(h \dashv u, D_2^{u,h})$, $MD(h \dashv u, D_1^{u,h})$, and $MD(h \dashv u, D_2^{u,h})$ equal 0 follow by symmetry.

From Lemma 2.21 we have $MD(h \dashv u, D_1^{u,h}) \geq 0$ and $MD(h \dashv u, D_2^{u,h}) = 0$. Hence, from our assumptions we have $CF(h \dashv u, D_1^{u,h}) \leq 0$ and $CF(h \dashv u, D_2^{u,h}) > 0$. From now on we assume $CF(h \dashv u, D_1^{u,h}) \cdot CF(h \dashv u, D_2^{u,h}) > -1$. So, the numerator of the fraction can be written as follows

$$\begin{aligned} MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) - MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= MB(h \dashv u, D_1^{u,h}) + MB(h \dashv u, D_2^{u,h})(1 - MB(h \dashv u, D_1^{u,h})) + \\ &\quad - MD(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_2^{u,h})(1 - MD(h \dashv u, D_1^{u,h})) = \\ &= MB(h \dashv u, D_2^{u,h}) - MD(h \dashv u, D_1^{u,h}) = \\ &= (MB(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_1^{u,h})) + \\ &\quad + (MB(h \dashv u, D_2^{u,h}) - MD(h \dashv u, D_2^{u,h})) \end{aligned}$$

From the observation that for $i = 1, 2$, we have $1 - \min\{MB(h \dashv u, D_i^{u,h}), MD(h \dashv u, D_i^{u,h})\} = 1$, it follows that

$$\begin{aligned} MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) - MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) &= \\ &= \frac{MB(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_1^{u,h})}{1 - \min\{MB(h \dashv u, D_1^{u,h}), MD(h \dashv u, D_1^{u,h})\}} + \\ &+ \frac{MB(h \dashv u, D_2^{u,h}) - MD(h \dashv u, D_2^{u,h})}{1 - \min\{MB(h \dashv u, D_2^{u,h}), MD(h \dashv u, D_2^{u,h})\}} = \\ &= CF(h \dashv u, D_1^{u,h}) + CF(h \dashv u, D_2^{u,h}) \end{aligned}$$

Note that $MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MB(h \dashv u, D_2^{u,h})$ and $MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = MD(h \dashv u, D_1^{u,h})$. We have that the denominator of the fraction equals

$$\begin{aligned} 1 - \min\{MB_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}), MD_{\parallel}(h \dashv u, D_1^{u,h} \parallel D_2^{u,h})\} &= \\ &= 1 - \min\{MB(h \dashv u, D_2^{u,h}), MD(h \dashv u, D_1^{u,h})\} \end{aligned}$$

It can easily be shown that

$$\begin{aligned} MB(h \dashv u, D_2^{u,h}) &= \frac{MB(h \dashv u, D_2^{u,h}) - MD(h \dashv u, D_2^{u,h})}{1 - \min\{MB(h \dashv u, D_2^{u,h}), MD(h \dashv u, D_2^{u,h})\}} = \\ &= |CF(h \dashv u, D_2^{u,h})| \end{aligned}$$

Furthermore, we can show that

$$\begin{aligned} MD(h \dashv u, D_1^{u,h}) &= - \frac{MB(h \dashv u, D_1^{u,h}) - MD(h \dashv u, D_1^{u,h})}{1 - \min\{MB(h \dashv u, D_1^{u,h}), MD(h \dashv u, D_1^{u,h})\}} = \\ &= |CF(h \dashv u, D_1^{u,h})| \end{aligned}$$

So, we have

$$CF'(h \dashv u, D_1^{u,h} \parallel D_2^{u,h}) = \frac{CF(h \dashv u, D_1^{u,h}) + CF(h \dashv u, D_2^{u,h})}{1 - \min\{|CF(h \dashv u, D_1^{u,h})|, |CF(h \dashv u, D_2^{u,h})|\}}$$

■

The Propositions 2.50, 2.51, 2.52 and 2.53 together yield the result stated in Theorem 2.54.

THEOREM 2.54. *Let the function CF' be defined according to Definition 2.28 and the function \overline{CF} according to Definition 2.30. Then, $CF' = \overline{CF}$.*

2.4.3. Summary of the Results

We have investigated the relations between the certainty factor functions CF , CF' and \overline{CF} . In Section 2.4.1 we have shown that the certainty factor function CF defined by Shortliffe and Buchanan and the function CF' actually used by them in implementations of the model, do not always render the same function values for the arguments of interest. In Section 2.4.2 we have shown that the approximation function \overline{CF} respects the definition of CF' . In fact, CF' and \overline{CF} coincide. In Figure 2.10, these results have been inserted into the diagram of functions introduced in Section 2.2.

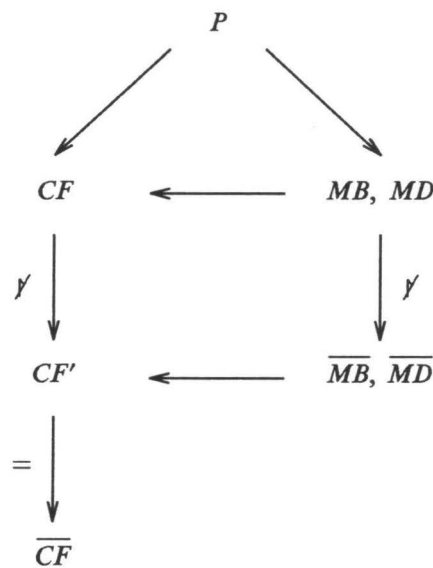


FIGURE 2.10. The diagram of functions.

The overall result of the preceding subsections shows that the certainty factor model is not correct with respect to the probabilistic foundation of the model as suggested by Shortliffe and Buchanan. The separate, detailed results we have discussed, identified conditions under which at least some components of

the model behave correctly. Recall that all results should be taken relative to the way we have introduced and handled the notion of derivation in the probabilistic foundation of the model.

We have not paid attention to the question whether it is to be expected that the identified conditions will be met in practice. Furthermore, although we have not analysed the impact of the application of the model in situations in which the conditions are not fulfilled, we feel, supported by the literature on the subject, that the model's behaviour degrades noticeably in such situations.

Shortliffe and Buchanan themselves have experimented with the model in the context of the MYCIN system using sampling data simulating several hundred patients, to compare the computed certainty factors, that is, the function values of \overline{CF} , with the correct probabilistic values, that is, the function values of CF , [SHOR84]. In this experiment, they have focussed on the combination function for co-concluding production rules. They observed that in most of the cases, the computed certainty factor does not differ radically from the theoretical probabilistic value. However, they have observed that the more the combination function for co-concluding production rules is applied for a given hypothesis, the more the computed values tend to deviate from the theoretical ones. Furthermore, their test showed that the most erroneous values arose from cases in which the different derivations of the hypothesis under consideration were strongly interrelated. We add to these observations that since the vast majority of the production rules of MYCIN contained positive certainty factors, the problematic case of conflicting derivations cannot have occurred very often. The experiment of Shortliffe and Buchanan therefore did not reflect the impact of conflicting evidence.

Guided by their experiment, Shortliffe and Buchanan themselves have warned against application of their model to other domains without due consideration. The certainty factor model however is incorporated as a special feature in many present-day, commercially available expert system shells. The model therefore is likely to be applied to any type of domain.

Chapter 3

Belief Networks

In the mid-eighties a new trend in reasoning with uncertainty in knowledge-based systems became discernable: several (mathematically correct) probabilistic models were proposed, each departing from a so-called *belief network*, see for example [SHAC86, PEAR88, SPIE86b]. Informally speaking, a belief 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 joint probability distribution on the variables. The phrase *belief network* has been adopted from J. Pearl, [PEAR88]. Several other phrases are used to denote the same concept: D.J. Spiegelhalter uses the phrase *causal graph* [SPIE86b], and the phrase *influence diagram* is used by R.D. Shachter [SHAC86]. Statisticians often use the phrase *recursive model* to denote similar graphical representations of a problem domain, see for example [WERM83, KIIV84].

This chapter presents a theoretical introduction to belief networks. In Section 3.1 some preliminaries are provided. Section 3.2 discusses the representation of a problem domain in a belief network in general. We present a mathematically detailed description of the notion of a belief network which, in the relevant literature often only introduced informally, is the common denominator of the work by Shachter, Pearl, Spiegelhalter and others. In Section 3.3 we briefly discuss some general properties a scheme for processing evidence in a belief network has to meet. It should be noted that although all network models proposed so far build on the same notion of belief network, they differ considerably in their schemes for processing evidence. In Section 3.4 one such scheme will be dealt with in some detail.

We note that as far as reasoning with uncertain information is concerned, the use of graphical representations of a problem domain is currently not only

studied from the perspective of probability theory, but from the perspective of Dempster-Shafer theory as well, see for example [SHEN86,DEMP88].

3.1. PRELIMINARIES: GRAPH THEORY AND PROBABILITY THEORY

In this section, we review several notions from graph theory that will play a central role in the remainder of this chapter. Furthermore, we once more provide some preliminaries concerning probability theory.

3.1.1. Graph Theory

We review some basic notions from graph theory. For further information the reader is referred to [WILS79,BERG73].

Generally, two types of graphs are discerned: undirected graphs and directed ones.

DEFINITION 3.1. *An undirected graph G is an ordered pair $G = (V(G), E(G))$, where $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, is a finite set of vertices and $E(G)$ is a family of unordered pairs (V_i, V_j) , $V_i, V_j \in V(G)$, called edges. Two vertices V_i and V_j are called adjacent or neighbouring vertices in G if $(V_i, V_j) \in E(G)$. The set of all neighbours of vertex V_i in G is denoted by $\nu_G(V_i)$.*

A directed graph (or digraph, for short) G is an ordered pair $G = (V(G), A(G))$, where $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, is a finite set of vertices and $A(G)$ is a family of ordered pairs (V_i, V_j) , $V_i, V_j \in V(G)$, called arcs. Vertex V_j is called a successor of vertex V_i if there is an arc $(V_i, V_j) \in A(G)$; the set of all successors of V_i in G is denoted by $\sigma_G(V_i)$. Similarly, vertex V_i is called a predecessor of vertex V_j if there is an arc $(V_i, V_j) \in A(G)$; the set of all predecessors of V_i in G is denoted by $\pi_G(V_i)$. The set of all neighbours of vertex V_i in G is defined as $\nu_G(V_i) = \sigma_G(V_i) \cup \pi_G(V_i)$.

In the sequel, we will often drop the subscript G from σ_G etc. as long as ambiguity cannot occur.

DEFINITION 3.2. *Let $G = (V(G), A(G))$ be a digraph. The underlying graph H of G is the undirected graph $H = (V(H), E(H))$ where $V(H) = V(G)$ and $E(H)$ is obtained from $A(G)$ by replacing each arc $(V_i, V_j) \in A(G)$ by the edge (V_i, V_j) .*

In the following Definitions 3.3 and 3.4 some notions are introduced concerning undirected graphs. These notions however can easily be extended to apply to directed graphs by taking the directions of the arcs into account.

DEFINITION 3.3. *An undirected graph $G = (V(G), E(G))$ is a simple graph if $E(G)$ is a set and $(V_i, V_i) \notin E(G)$ for all $V_i \in V(G)$.*

In the sequel, we take all (directed and undirected) graphs to be simple.

DEFINITION 3.4. Let $G = (V(G), E(G))$ be an undirected graph. A path from V_0 to V_k , $V_0, V_k \in V(G)$, in G is a sequence of vertices V_0, V_1, \dots, V_k such that $(V_{i-1}, V_i) \in E(G)$, $i = 1, \dots, k$, $k \geq 0$; k is called the length of the path.

If for each pair of vertices $V_i, V_j \in V(G)$ there is a path from V_i to V_j in G , then G is called a connected graph; otherwise G is disconnected.

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 or shortcut of an elementary cycle $V_0, V_1, \dots, V_k = V_0$ is an edge (V_i, V_j) , $i \neq (j \pm 1) \bmod (k+1)$.

G is called a cyclic graph if it contains at least one cycle; a graph without any cycles is called acyclic.

We conclude this subsection with two more definitions.

DEFINITION 3.5. Let $G = (V(G), E(G))$ be an undirected graph. The order of G is the number of vertices in G . The size of G is the number of edges in G . G is a complete n -graph, $n \geq 1$, if it has order n and size $\binom{n}{2}$, that is, a graph is complete if there exists an edge between each pair of distinct vertices.

An undirected graph $H = (V(H), E(H))$ is a subgraph of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. A subgraph H of G is a full subgraph of G if $E(H) = E(G) \cap (V(H) \times V(H))$; we say that the full subgraph H is induced by $V(H)$.

A clique in G is a full subgraph H of G which is complete. H is called a maximal clique if there does not exist a clique H' in G differing from H such that H is a full subgraph of H' . The set of all maximal cliques in G is called the clique set of G and is denoted by $Cl(G)$.

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

DEFINITION 3.6. A forest is an undirected graph which is acyclic. A tree is a connected forest.

Let $G = (V(G), E(G))$ be a connected undirected graph. Furthermore, let $T = (V(T), E(T))$ with $V(T) = V(G)$ and $E(T) \subseteq E(G)$ be a subgraph of G such that T is a tree. Then, T is called a spanning tree of G .

3.1.2. Probability Theory Revisited

In this subsection, again some preliminaries concerning probability theory are provided, this time not departing from a set-theoretic point of view, but from an algebraic one.

In an expert system, knowledge concerning the problem domain usually is represented in a special knowledge-representation formalism such as for example the production-rule formalism we have encountered in Chapter 2. In the present chapter we do not consider such knowledge-representation schemes

nor do we discuss the reasoning methods associated with these formalisms. Here, we assume that knowledge is simply represented in statistical variables and their probabilistic interrelationships. We assume that these variables can only take one of two values, thus allowing to view them as logical, propositional variables. The generalization to variables with discrete multiple values, however, is rather straightforward.

In the following definition the notion of a Boolean algebra of propositions is introduced; for further information the reader is referred to [BIRK77].

DEFINITION 3.7. *A Boolean algebra \mathcal{B} is a set of elements with two binary operations \wedge (conjunction) and \vee (disjunction), a unary operation \neg (negation) and two constants false and true which (by equality according to logical truth tables) adhere to the usual axioms.*

On a Boolean algebra \mathcal{B} we define a partial order \leq as follows: for any $x_1, x_2 \in \mathcal{B}$, we say that $x_1 \leq x_2$ if $x_2 = x_1 \vee x_2$ or (equivalently) if $x_1 = x_1 \wedge x_2$.

A subset of elements $\mathcal{G} = \{g_1, \dots, g_n\}$, $n \geq 1$, of a Boolean algebra \mathcal{B} is said to be a set of generators for \mathcal{B} if each element of \mathcal{B} can be represented in terms of the elements $g_i \in \mathcal{G}$, $i = 1, \dots, n$, and the operations \wedge , \vee and \neg . A set of generators \mathcal{G} for \mathcal{B} is said to be free if every mapping of elements of \mathcal{G} into an arbitrary Boolean algebra \mathcal{B}' can be extended to a homomorphism of \mathcal{B} into \mathcal{B}' .

A Boolean algebra \mathcal{B} is free if it has a finite set $\mathcal{A} = \{a_1, \dots, a_n\}$, $n \geq 1$, of free generators; we say that \mathcal{B} is (finitely) generated by \mathcal{A} . We use $\mathcal{B}(a_1, \dots, a_n)$ to denote the free Boolean algebra \mathcal{B} generated by \mathcal{A} ; from now on, we will refer to \mathcal{A} as the set of atomic propositions and to \mathcal{B} as the Boolean algebra of propositions.

It is well-known that a free Boolean algebra with n free generators has 2^{2^n} elements, $n \geq 1$. In the sequel, for any subset $\{a_{i_1}, \dots, a_{i_k}\} \subseteq \{a_1, \dots, a_n\}$, $1 \leq k \leq n$, we simply use $\mathcal{B}(a_{i_1}, \dots, a_{i_k})$ to denote the subalgebra of $\mathcal{B}(a_1, \dots, a_n)$ generated by $\{a_{i_1}, \dots, a_{i_k}\}$.

DEFINITION 3.8. *Let $\mathcal{A} = \{a_1, \dots, a_n\}$, $n \geq 1$, be a set of atomic propositions and let $\mathcal{B}(a_1, \dots, a_n)$ be the free Boolean algebra generated by \mathcal{A} . For $i = 1, \dots, m$, $m \geq 0$, let A_i be a variable over $\mathcal{B}(a_1, \dots, a_n)$. A Boolean polynomial function in the variables A_1, \dots, A_m is a function $F: \mathcal{B}(a_1, \dots, a_n)^m \rightarrow \mathcal{B}(a_1, \dots, a_n)$. We use $\mathcal{B}(a_1, \dots, a_n)[A_1, \dots, A_m]$ to denote the set of Boolean polynomial functions in A_1, \dots, A_m .*

In the following definition we introduce a so-called configuration function as a special type of Boolean polynomial function.

DEFINITION 3.9. Let $\mathcal{A} = \{a_1, \dots, a_n\}$, $n \geq 1$, be a set of atomic propositions and let $\mathcal{B}(a_1, \dots, a_n)$ be the free Boolean algebra generated by \mathcal{A} . Let $A = \{A_{i_1}, \dots, A_{i_k}\}$, $0 \leq k \leq n$, be a set of variables over $\mathcal{B}(a_1, \dots, a_n)$. Now, let $F_A \in \mathcal{B}(a_1, \dots, a_n)[A_{i_1}, \dots, A_{i_k}]$ be the Boolean polynomial function defined by $F_A = \text{true}$ if $k = 0$ and $F_A(A_{i_1}, \dots, A_{i_k}) = A_{i_1} \wedge \dots \wedge A_{i_k}$ otherwise. Let $B_i = \{a_i, \neg a_i\}$, $i = 1, \dots, n$. We define the configuration function C_A as the restriction of F_A to $B_{i_1} \times \dots \times B_{i_k}$, that is, $C_A = F_A|_{B_{i_1} \times \dots \times B_{i_k}}$. A function value c_A of C_A is called a configuration of A .

In the sequel, we will often use the notation $\{c_A\}$ to denote the set of all configurations of the set of variables A .

EXAMPLE 3.10. Consider the Boolean algebra of propositions $\mathcal{B}(a_1, \dots, a_8)$. Let $A = \{A_1, A_3, A_7\}$ be a set of variables over $\mathcal{B}(a_1, \dots, a_8)$. Then, $F_A(A_1, A_3, A_7) = A_1 \wedge A_3 \wedge A_7$. The configuration function C_A will be viewed as defined by $C_A(A_1, A_3, A_7) = A_1 \wedge A_3 \wedge A_7$ where A_i now is taken to be a variable over $\{a_i, \neg a_i\}$, $i = 1, 3, 7$, only. The conjunction $\neg a_1 \wedge a_3 \wedge a_7 \in \mathcal{B}(a_1, \dots, a_8)$ is an example of a configuration of A . Note that a configuration is a semantical notion whereas a configuration function is a syntactic one. ■

We introduce the notion of a probability distribution on a Boolean algebra of propositions.

DEFINITION 3.11. Let \mathcal{B} be a Boolean algebra of propositions defined according to Definition 3.7. Let Pr be a function $Pr: \mathcal{B} \rightarrow [0, 1]$ such that

- (1) Pr is positive, that is, for all $x \in \mathcal{B}$, we have $Pr(x) \geq 0$, and furthermore $Pr(\text{false}) = 0$,
- (2) Pr is normed, that is, we have $Pr(\text{true}) = 1$, and
- (3) Pr is additive, that is, for all $x_1, x_2 \in \mathcal{B}$, if $x_1 \wedge x_2 = \text{false}$ then $Pr(x_1 \vee x_2) = Pr(x_1) + Pr(x_2)$.

Then, Pr is called a probability distribution on \mathcal{B} . The pair (\mathcal{B}, Pr) is called a probability algebra.

Let $\mathcal{B}(a_1, \dots, a_n)$ be a Boolean algebra of propositions and let $A = \{A_1, \dots, A_n\}$ be a set of variables over $\mathcal{B}(a_1, \dots, a_n)$. In this chapter we will frequently exploit the property that a probability distribution Pr on a Boolean algebra $\mathcal{B}(a_1, \dots, a_n)$ is uniquely defined by its values $Pr(c_A)$ for each configuration c_A of A ; this property will be proven formally in Chapter 4. Note that since there are 2^n possible configurations of A , to explicitly represent Pr in a straightforward manner would require 2^n probabilities. We shall see in Section 3.2, however, that in some cases far less probabilities suffice for uniquely representing Pr .

Recall that in Section 2.2.1 we associated probabilities with sets instead of with logical propositions. However, it can easily be shown that the probability of an event is equivalent to the probability of the truth of the proposition asserting the occurrence of the event. The following proposition states this well-known equivalence more formally. We provide the proposition and part of its proof merely because it conveys many ideas that we will return to in the next chapter. For further details, the reader is referred to [FINE70].

PROPOSITION 3.12. *Let \mathcal{A} be a set of atomic propositions and let \mathcal{B} be the Boolean algebra of propositions generated by \mathcal{A} as defined in Definition 3.7. Let Pr be a probability distribution on \mathcal{B} . Then, there exists a sample space Ω , a probability function P on Ω , and an isomorphism $\iota: \mathcal{B} \rightarrow \mathcal{F}$ where \mathcal{F} is the set of subsets of Ω , such that*

- (1) *for all $x_1, x_2 \in \mathcal{B}$, we have $\iota(x_1 \wedge x_2) = \iota(x_1) \cap \iota(x_2)$,*
- (2) *for all $x_1, x_2 \in \mathcal{B}$, we have $\iota(x_1 \vee x_2) = \iota(x_1) \cup \iota(x_2)$,*
- (3) *for all $x \in \mathcal{B}$, we have $\iota(\neg x) = \overline{\iota(x)}$,*
- (4) *\mathcal{F} equals the free Boolean algebra generated by $\{\iota(x) \mid x \in \mathcal{A}\}$, and*
- (5) *for each $x \in \mathcal{B}$, we have $P(\iota(x)) = Pr(x)$.*

We have that P is uniquely defined by Pr . Furthermore, the algebras (\mathcal{F}, P) and (\mathcal{B}, Pr) are isomorphic.

PROOF. We only provide a sketch of the proof. Let $\mathcal{A} = \{a_1, \dots, a_n\}$, $n \geq 1$. Now, let $\Omega = \{\bigwedge_{i=1}^n A_i \mid A_i = a_i \text{ or } A_i = \neg a_i, a_i \in \mathcal{A}\}$. Note that the elements of Ω are all configurations of length n in which for each $i = 1, \dots, n$, either a_i or $\neg a_i$ occurs. It will be evident that Ω has 2^n elements. In the remainder of this proof the elements of Ω are enumerated as $\omega_1, \dots, \omega_{2^n}$.

Using De Morgan's laws and the distributive laws, any element of \mathcal{B} can be represented as a disjunction of elements of Ω : for each $x \in \mathcal{B}$ there exists a unique set of indices $\mathcal{J}_x \subseteq \{1, \dots, 2^n\}$ such that $x = \bigvee_{i \in \mathcal{J}_x} \omega_i$ where $\omega_i \in \Omega$.

When x is represented as $\bigvee_{i \in \mathcal{J}_x} \omega_i$, we say that x is in *disjunctive normal form*.

We define a mapping ι as follows: for $x = \bigvee_{i \in \mathcal{J}_x} \omega_i$ where $\mathcal{J}_x \subseteq \{1, \dots, 2^n\}$, we take $\iota(x) = \{\omega_i \mid i \in \mathcal{J}_x\}$. Note that ι is well-defined since for a given x the set \mathcal{J}_x is unique.

It is obvious that we have $\iota(\text{false}) = \emptyset$ and $\iota(\text{true}) = \Omega$. Furthermore, we have the following properties of this mapping ι :

- (1) *for all $x_1, x_2 \in \mathcal{B}$, $\iota(x_1 \wedge x_2) = \iota(x_1) \cap \iota(x_2)$.*

Suppose that we have $x_1 = \bigvee_{i_1 \in \mathcal{J}_1} \omega_{i_1}$ and $x_2 = \bigvee_{i_2 \in \mathcal{J}_2} \omega_{i_2}$, where

$\mathcal{J}_1, \mathcal{J}_2 \subseteq \{1, \dots, 2^n\}$. So, $x_1 \wedge x_2 = (\bigvee_{i_1 \in \mathcal{J}_1} \omega_{i_1}) \wedge (\bigvee_{i_2 \in \mathcal{J}_2} \omega_{i_2})$. Using the distributive laws, $x_1 \wedge x_2$ can be written as $\bigvee_{i_1 \in \mathcal{J}_1, i_2 \in \mathcal{J}_2} (\omega_{i_1} \wedge \omega_{i_2})$.

From our definition of Ω , we have that $\omega_{i_1} \wedge \omega_{i_2} = \text{false}$ for $i_1 \neq i_2$. It follows that $x_1 \wedge x_2 = \bigvee_{i \in \mathcal{J}_1 \cap \mathcal{J}_2} \omega_i$. Consequently, we have

$$\begin{aligned} \iota(x_1 \wedge x_2) &= \{\omega_i \mid i \in \mathcal{J}_1 \cap \mathcal{J}_2\} = \{\omega_{i_1} \mid i_1 \in \mathcal{J}_1\} \cap \{\omega_{i_2} \mid i_2 \in \mathcal{J}_2\} = \\ &= \iota(x_1) \cap \iota(x_2) \end{aligned}$$

$$(2) \quad \text{for all } x_1, x_2 \in \mathcal{B}, \iota(x_1 \vee x_2) = \iota(x_1) \cup \iota(x_2).$$

$$(3) \quad \text{for all } x \in \mathcal{B}, \iota(\neg x) = \overline{\iota(x)}.$$

It will be evident that the mapping ι is an isomorphism, since it has an inverse mapping ι^{-1} .

The algebra \mathcal{F} obviously equals $\{\iota(x) \mid x \in \mathcal{B}\}$. We have that \mathcal{F} is the free Boolean algebra generated by $\{\iota(x) \mid x \in \mathcal{B}\}$.

From the properties of the probability distribution Pr on \mathcal{B} we have that $P (= Pr \circ \iota^{-1})$ is additive and $[0,1]$ -valued on \mathcal{F} and therefore is a probability function on \mathcal{F} . It can easily be shown that (\mathcal{B}, Pr) and (\mathcal{F}, P) are isomorphic. ■

From the previous proposition we have that a probability distribution Pr on a Boolean algebra of propositions \mathcal{B} has the usual properties. Lemma 3.13 repeats some convenient ones.

LEMMA 3.13. *Let (\mathcal{B}, Pr) be a probability algebra as defined in Definition 3.11. Then,*

- (1) *for all $x \in \mathcal{B}$, we have $Pr(x) + Pr(\neg x) = 1$,*
- (2) *for all $x_1, x_2 \in \mathcal{B}$, $Pr(x_1 \vee x_2) + Pr(x_1 \wedge x_2) = Pr(x_1) + Pr(x_2)$, and*
- (3) *for all $x_1, x_2 \in \mathcal{B}$, if $x_1 \leq x_2$ then $Pr(x_1) \leq Pr(x_2)$.*

The following lemma is rather straightforward.

LEMMA 3.14. *Let $\mathcal{B}(a_1, \dots, a_n)$, $n \geq 1$, be a Boolean algebra of propositions defined according to Definition 3.7. Let Pr be a probability distribution on $\mathcal{B}(a_1, \dots, a_n)$. Then, for each i , $1 \leq i \leq n$, we have that the probabilities*

$$Pr(x) = Pr(x \wedge a_i) + Pr(x \wedge \neg a_i)$$

for all $x \in \mathcal{B}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$, define a probability distribution on $\mathcal{B}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$.

PROOF. The lemma follows from the observation that the properties mentioned in Definition 3.11 hold. ■

The probability distribution defined by the probabilities $Pr(x) = P(x \wedge a_i) + Pr(x \wedge \neg a_i)$ as indicated in the preceding lemma is called the *marginal distribution* on $\mathcal{B}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$. The 'entire' probability distribution Pr is often called the *joint probability distribution* to discern it explicitly from marginal distributions derived from it. Note that by recursively applying Lemma 3.14 we may obtain a marginal distribution on any subalgebra of $\mathcal{B}(a_1, \dots, a_n)$.

In the sequel, we will often have to deal with the situation that several pieces of evidence become available which should be taken into account in future probabilistic statements. For this purpose, we introduce the notion of a conditional probability.

DEFINITION 3.15. Let (\mathcal{B}, Pr) be a probability algebra defined according to Definition 3.11. For each $x, y \in \mathcal{B}$ with $Pr(y) > 0$, the conditional probability of x given y , denoted as $Pr(x | y)$, is defined as

$$Pr(x | y) = \frac{Pr(x \wedge y)}{Pr(y)}$$

In the sequel, we will implicitly assume that the conditional probabilities we specify are defined unless explicitly stated otherwise.

The following lemma states that given a specific piece of evidence we may compute a revised probability distribution.

LEMMA 3.16. Let $\mathcal{B}(a_1, \dots, a_n)$, $n \geq 1$, be a Boolean algebra of propositions defined according to Definition 3.7. Let Pr be a joint probability distribution on $\mathcal{B}(a_1, \dots, a_n)$. Then, for a given $e \in \{a_i, \neg a_i\}$, $1 \leq i \leq n$, the conditional probabilities $Pr(x | e)$ for all $x \in \mathcal{B}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$ define a probability distribution on $\mathcal{B}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$.

The probability distribution defined by the conditional probabilities $Pr(x | e)$ as in the preceding lemma is called the *updated probability distribution* given e ; in the sequel, this probability distribution will often be written as Pr^e . Furthermore, we will use the phrase to *update* a probability distribution to denote the process of computing the updated probability distribution given some piece of evidence. For updating a joint probability distribution for successively obtained evidence the preceding lemma may be applied recursively: let $e_{i_j} \in \{a_{i_j}, \neg a_{i_j}\}$, $j = 1, \dots, k$, $1 \leq k \leq n$, $n \geq 1$, and let $Pr^{e_1, \dots, e_{i_j}}$ be the updated probability distribution given e_1, \dots, e_{i_j} . Then, we have that $Pr^{e_1, \dots, e_{i_k}}(x) = Pr^{e_1, \dots, e_{i_{k-1}}}(x | e_{i_k})$, for all $x \in \mathcal{B}(a_{i_{k+1}}, \dots, a_{i_n})$.

In the sequel, we take Bayes' Theorem introduced in Chapter 2 to accord with the algebraic point of view. The following theorem is known as the *chain rule*.

THEOREM 3.17. *Let $\mathcal{B}(a_1, \dots, a_n)$, $n \geq 1$, be a Boolean algebra of propositions as defined in Definition 3.7. Let Pr be a joint probability distribution on $\mathcal{B}(a_1, \dots, a_n)$. Then, for all $x_i \in \{a_i, \neg a_i\}$, $i = 1, \dots, n$, we have*

$$Pr(x_1 \wedge \dots \wedge x_n) = Pr(x_n | x_1 \wedge \dots \wedge x_{n-1}) \dots Pr(x_2 | x_1) \cdot Pr(x_1)$$

Note that using configuration functions, the property from the previous theorem may be rewritten as

$$\begin{aligned} Pr(A_1 \wedge \dots \wedge A_n) &= \\ &= Pr(A_n | A_1 \wedge \dots \wedge A_{n-1}) \dots Pr(A_2 | A_1) \cdot Pr(A_1) \end{aligned}$$

where each A_i is a variable taking values from $\{a_i, \neg a_i\}$, $i = 1, \dots, n$. From now on we will adhere to this point of view and take a variable A_i to be a variable over $B_i = \{a_i, \neg a_i\}$.

We conclude this subsection with one more definition.

DEFINITION 3.18. *Let $\mathcal{B}(a_1, \dots, a_n)$, $n \geq 1$, be a Boolean algebra of propositions as defined in Definition 3.7. Let Pr be a joint probability distribution on $\mathcal{B}(a_1, \dots, a_n)$. Let $A = \{A_1, \dots, A_n\}$ be a set of variables where each A_i is a variable over $B_i = \{a_i, \neg a_i\}$, $i = 1, \dots, n$. Furthermore, let $X, Y, Z \subseteq A$ and let C_X , C_Y and C_Z be configuration functions for the sets X , Y and Z , respectively, as defined in Definition 3.9. The set of variables X is said to be conditionally independent of Y given Z , denoted as $I_{Pr}(X, Z, Y)$, if $Pr(C_X | C_Y \wedge C_Z) = Pr(C_X | C_Z)$.*

The following lemma can now easily be proven.

LEMMA 3.19. *Let $\mathcal{B}(a_1, \dots, a_n)$, $n \geq 1$, be a free Boolean algebra of propositions. Let Pr be a joint probability distribution on $\mathcal{B}(a_1, \dots, a_n)$. Let $A = \{A_1, \dots, A_n\}$ be a set of variables where each A_i is a variable over $B_i = \{a_i, \neg a_i\}$, $i = 1, \dots, n$. Let $X, Y, Z, W \subseteq A$. Furthermore, let the relation I_{Pr} have the meaning as in Definition 3.18. Then, the following properties hold:*

- (1) $I_{Pr}(X, Z, Y)$ if and only if $I_{Pr}(Y, Z, X)$.
- (2) If $I_{Pr}(X, Z, Y \cup W)$ then $I_{Pr}(X, Z, Y)$ and $I_{Pr}(X, Z, W)$.
- (3) If $I_{Pr}(X, Z, Y \cup W)$ then $I_{Pr}(X, Z \cup W, Y)$.
- (4) If $I_{Pr}(X, Z, Y)$ and $I_{Pr}(X, Z \cup Y, W)$ then $I_{Pr}(X, Z, Y \cup W)$.

The first property mentioned in the preceding lemma is called the property of *symmetry*. The second one is the *decomposition* property. The third property is the property of *weak union* and the fourth one is called the *contraction* property. For an in-depth discussion of the independence relation I_{Pr} , the reader is referred to [PEAR88].

3.2. KNOWLEDGE REPRESENTATION IN A BELIEF NETWORK

In Section 3.1.2 we have remarked that in this chapter we do not depart from the knowledge representation schemes generally employed in knowledge-based systems: we assume that knowledge is represented in statistical variables (viewed as propositional variables) and their probabilistic interrelationships. In this section we discuss this representation scheme in further detail.

Belief networks provide a formalism for representing a problem domain. A belief network comprises two parts: a *qualitative representation* of the problem domain and an associated *quantitative representation*. The qualitative part of a belief network takes the form of an acyclic directed graph $G = (V(G), A(G))$ with vertices $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and arcs $A(G)$. Each vertex V_i in $V(G)$ represents a statistical variable that can take one of a set of values. In the sequel, we assume that the statistical variables can take only one of the truth values *true* and *false*. We will adhere to the following notational convention: v_i denotes the proposition that the variable V_i takes the truth value *true*; $V_i = \text{false}$ will be denoted by $\neg v_i$. Informally speaking, we take an arc $(V_i, V_j) \in A(G)$ to represent a direct 'influential' or 'causal' relationship between the linked 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 the sequel, we take the digraph to be configured by an expert from human judgment; hence the phrase *belief network*.

Associated with the graphical part of a belief network is a numerical assessment of the 'strengths' of the represented relationships: with each vertex is associated a set of (conditional) probabilities which describe the influence of the values of the predecessors of the vertex on the values of the vertex itself.

We define the notion of a belief network more formally.

DEFINITION 3.20. *A belief network is a tuple $B = (G, \Gamma)$ such that*

- (1) $G = (V(G), A(G))$ is an acyclic directed graph with vertices $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and arcs $A(G)$, and
- (2) $\Gamma = \{\gamma_{V_i} \mid V_i \in V(G)\}$ is a set of real-valued nonnegative functions $\gamma_{V_i}: \{v_i, \neg v_i\} \times \{c_{\pi(V_i)}\} \rightarrow [0, 1]$, called (conditional probability) assessment functions, such that for each configuration $c_{\pi(V_i)}$ of $\pi(V_i)$, we have $\gamma_{V_i}(\neg v_i \mid c_{\pi(V_i)}) = 1 - \gamma_{V_i}(v_i \mid c_{\pi(V_i)})$, $i = 1, \dots, n$.

Note that in the previous definition V_i is viewed as a vertex from the graph and as a variable over $\{v_i, \neg v_i\}$, alternatively. For ease of exposition, we assume in the remainder of this thesis that the graphical part of a belief network is connected; our observations, however, can easily be extended to apply to disconnected acyclic digraphs.

In order to draw a link between the qualitative and quantitative parts of a belief network, we assign a probabilistic meaning to the topology of the digraph G of the network. For each vertex $V_i \in V(G)$, we define the set $\alpha(V_i)$

of (strictly) anterior vertices of V_i by $\alpha(V_i) = \{V_j \mid V_j \in V(G) \setminus \pi(V_i) \text{ and there is no path from } V_i \text{ to } V_j\}$. Informally speaking, the digraph now is taken to represent the following independency relationships: each variable $V_i \in V(G)$ is conditionally independent of the variables from $\alpha(V_i)$ given its parent variables $\pi(V_i)$. We say that a joint probability distribution Pr respects the independency relationships portrayed by G if for each variable $V_i \in V(G)$ we have that $Pr(V_i \mid C_{\alpha(V_i)} \wedge C_{\pi(V_i)}) = Pr(V_i \mid C_{\pi(V_i)})$. Since in the sequel we will primarily be concerned with undirected graphs, we do not discuss the link between probability distributions and directed graphs any further; for details, the reader is referred to [LAUR88b, PEAR88].

The following proposition states that the initial assessment functions of a belief network provide all information necessary for uniquely defining a joint probability distribution on the variables discerned that respects the independency relationships portrayed by the graphical part of the network.

PROPOSITION 3.21. *Let $B = (G, \Gamma)$ be a belief network as defined in the preceding definition, where $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra of propositions generated by $\{v_i \mid V_i \in V(G)\}$. Then,*

$$Pr(C_{V(G)}) = \prod_{V_i \in V(G)} \gamma_{V_i}(V_i \mid C_{\pi(V_i)})$$

defines a joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_n)$ that respects the independency relationships from G .

PROOF. A digraph without directed cycles allows at least one total ordering of its vertices such that any successor of a vertex in the graph follows it in the ordering. It follows that there is an ordering of the statistical variables such that in applying the chain rule each variable is conditioned only on the variables preceding it in the ordering. Choosing an appropriate ordering of $V(G)$, the conditional independency relationships portrayed by G can be exploited. By taking $Pr(v_i \mid c_{\pi(V_i)}) = \gamma_{V_i}(v_i \mid c_{\pi(V_i)})$ for each $V_i \in V(G)$ and all configurations $c_{\pi(V_i)}$ of $\pi(V_i)$, the property stated in the proposition follows immediately. For further details, see [KIIV84]. ■

Example 3.22 illustrates the notion of a belief network. Whenever possible, this example will be used as the running example; it has been taken from [LAUR88a].

EXAMPLE 3.22. Let $G = (V(G), A(G))$ with $V(G) = \{V_1, \dots, V_8\}$ be the acyclic digraph shown in Figure 3.1. We assume that this graph has been configured by an expert who for example observed that the value of the variable V_2 is only dependent directly upon the value of the variable V_1 . Let $\mathcal{B}(v_1, \dots, v_8)$ be the Boolean algebra of propositions associated with G as indicated in Proposition 3.21. Corresponding with this digraph G the expert

has assessed the following eighteen function values of the conditional probability assessment functions $\gamma_{V_1}, \dots, \gamma_{V_8}$:

$$\gamma_{V_1}(v_1)$$

$$\gamma_{V_2}(v_2 | v_1) \text{ and } \gamma_{V_2}(v_2 | \neg v_1)$$

$$\gamma_{V_3}(v_3)$$

$$\gamma_{V_4}(v_4 | v_3) \text{ and } \gamma_{V_4}(v_4 | \neg v_3)$$

$$\gamma_{V_5}(v_5 | v_3) \text{ and } \gamma_{V_5}(v_5 | \neg v_3)$$

$$\gamma_{V_6}(v_6 | v_2 \wedge v_4), \gamma_{V_6}(v_6 | v_2 \wedge \neg v_4), \gamma_{V_6}(v_6 | \neg v_2 \wedge v_4) \text{ and}$$

$$\gamma_{V_6}(v_6 | \neg v_2 \wedge \neg v_4)$$

$$\gamma_{V_7}(v_7 | v_5 \wedge v_6), \gamma_{V_7}(v_7 | v_5 \wedge \neg v_6), \gamma_{V_7}(v_7 | \neg v_5 \wedge v_6) \text{ and}$$

$$\gamma_{V_7}(v_7 | \neg v_5 \wedge \neg v_6)$$

$$\gamma_{V_8}(v_8 | v_6) \text{ and } \gamma_{V_8}(v_8 | \neg v_6)$$

Note that from these function values we can uniquely compute the remaining function values using $\gamma_{V_i}(\neg v_i | c_{\pi(V_i)}) = 1 - \gamma_{V_i}(v_i | c_{\pi(V_i)})$, $i = 1, \dots, 8$.

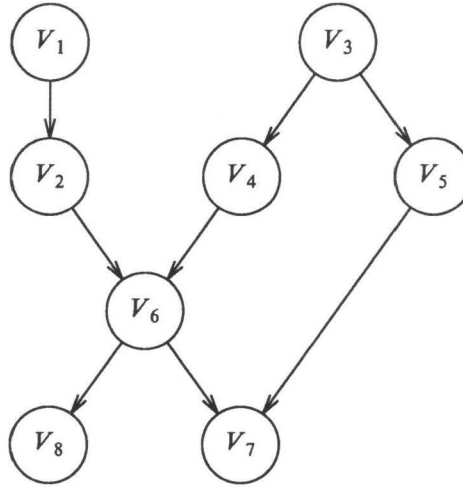


FIGURE 3.1. An acyclic digraph G .

The graph is taken to represent conditional independency relationships among the statistical variables V_1, \dots, V_8 . The graph for example shows that the variable V_7 is conditionally independent of V_3 and V_4 given V_5 and V_6 , that is, for any joint probability distribution P on $\mathcal{B}(v_1, \dots, v_8)$ respecting the independency relationships portrayed by G , we have that $P(V_7 | V_3 \wedge V_4 \wedge V_5 \wedge V_6) = P(V_7 | V_5 \wedge V_6)$. Exploiting these independency relationships, the joint probability distribution P can be expressed as the following product:

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

From Proposition 3.21 we have that the values of the assessment functions γ_{V_i} taken as (conditional) probabilities together define a specific joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_8)$: we therefore have that

$$\begin{aligned}
 Pr(V_1 \wedge \dots \wedge V_8) &= \\
 &= \gamma_{V_8}(V_8 | V_6) \cdot \gamma_{V_7}(V_7 | V_5 \wedge V_6) \cdot \gamma_{V_6}(V_6 | V_2 \wedge V_4) \cdot \gamma_{V_5}(V_5 | V_3) \cdot \\
 &\quad \cdot \gamma_{V_4}(V_4 | V_3) \cdot \gamma_{V_3}(V_3) \cdot \gamma_{V_2}(V_2 | V_1) \cdot \gamma_{V_1}(V_1)
 \end{aligned}$$

Any actual probability $Pr(c_{V(G)})$ can now be obtained by ‘filling in’ values for the statistical variables V_1 up to V_8 inclusive and then computing the resulting product on the right-hand side from the initially assessed probabilities. Note that in this example only eighteen probabilities suffice for uniquely representing a joint probability distribution on a Boolean algebra with eight free generators. ■

The representation of uncertainty in factors which are local to expressions giving a qualitative description of the domain, resembles the approach followed in the quasi-probabilistic models for dealing with uncertainty in rule-based systems in which the production rules constitute the qualitative representation of the domain.

3.3. EVIDENCE PROPAGATION IN A BELIEF NETWORK

In the preceding section we have introduced the notion of a belief network as a means for representing a problem domain, or, to be more precise, for representing a joint probability distribution. A belief network may be used for reasoning with uncertainty. For making probabilistic statements concerning the statistical variables discerned in the problem domain, we have to associate with a belief network two methods:

- (1) a method for (efficiently) computing probabilities of interest from the belief network, and
- (2) a method for processing evidence, that is, a method for entering evidence into the network and subsequently (efficiently) computing the updated probability distribution given the evidence. This process is generally called *evidence propagation*.

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

Recall that the assessment functions initially given for a belief network uniquely define a joint probability distribution Pr on 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 can therefore be computed from these 'local' function values. Calculation of an updated probability from the initially given joint probability distribution Pr in a straightforward manner, however, will generally not be restricted to performing computations which are local in terms of the graphical part of the network, and will become prohibitive for larger networks. For example, computing a conditional probability $Pr(\neg v_i | v_j)$ using Definition 3.15 would entail dividing two marginal probabilities each of which is the sum of an exponentially large number of probabilities computed from Γ as indicated in Example 3.22.

In the literature therefore, several less naive schemes for updating a joint probability distribution as evidence becomes available have been proposed. Although all methods proposed build on the same notion of a belief network, they differ considerably in concept and in computational complexity; as far as computational complexity is concerned, it should be noted that in the general case exact probabilistic inference in belief networks without any restrictions is NP-hard, [COOP87]. All proposals however have two important characteristics in common:

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

We briefly review some of the proposed schemes for evidence propagation.

R.D. Shachter has presented a method for propagating the impact of

evidence concerning a specified set of variables to a set of variables of interest. The general idea of his method is to eliminate vertices from the original graphical part of the belief network without changing the (updated) joint probability distribution; the topology of the graph is modified using a sequence of arc reversals, and vertex removals and additions, [SHAC86]. For each successive propagation of evidence again such a sequence of graph modifications has to be performed. The problem of optimizing a sequence of graph modifications has been further investigated, see for example [TRUN88].

The method for processing evidence presented by J.H. Kim and J. Pearl in [KIM83] is only applicable to singly connected digraphs, a restricted type of acyclic digraph. Their method leaves the original graphical representation of the problem domain unchanged. Updating the joint probability distribution after a piece of evidence has become available essentially entails each statistical variable (that is, each vertex) updating the joint probability distribution locally from messages it receives from its neighbours in the digraph, that is, from its predecessors as well as its successors, and then in turn sending new, updated messages to them. In his more recent work, [PEAR88], Pearl proposes additional methods for coping with (undirected) cycles.

S.L. Lauritzen and D.J. Spiegelhalter have presented another, elegant method for evidence propagation, [SPIE86b, LAUR88a]. They have observed that updating the joint probability distribution 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 belief network is not suitable as an architecture for propagating evidence directly. This observation, among other ones, motivated an initial transformation of the belief network into an undirected graphical and probabilistic representation of the problem domain. 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. For this purpose, Lauritzen and Spiegelhalter make use of the existing statistical theory of Markov random fields, see for example [PITM76, DARR80].

The work of Lauritzen and Spiegelhalter will be treated in further detail in Section 3.4; this method will play an important role in the remainder of the present thesis.

3.4. EVIDENCE PROPAGATION BY LAURITZEN AND SPIEGELHALTER

The method for evidence propagation presented by S.L. Lauritzen and D.J. Spiegelhalter departs from an *undirected* (graphical and probabilistic) representation of the problem domain. The method has been inspired by the existing statistical theory of Markov random fields, or more in specific, by the theory of *graphical models* (i.e. probabilistic models that can be represented by an undirected graph) in contingency tables. To be able to exploit this theory, the original *directed* belief network is transformed into an *undirected* so-called decomposable belief network which again consists of a qualitative representation of the problem domain, this time a so-called decomposable

graph, and a quantitative representation, now being a set of marginal distributions associated with the cliques of this graph.

This transformation scheme will be discussed in detail in Section 3.4.2. First, however, we will address in Section 3.4.1 the issue of assigning a probabilistic meaning to the topology of an undirected graph. We conclude this chapter with a discussion of the method for evidence propagation of Lauritzen and Spiegelhalter in Section 3.4.3.

3.4.1. Probabilistic Interpretation of the Topology of an Undirected Graph

In this subsection, we will discuss the relationship between probability distributions and undirected graphs. For this purpose, we introduce some notions from the theory of Markov random fields on finite graphs. For further information, the reader is referred to [PITM76]; [PEAR88] addresses the subject from the perspective of belief networks. In our discussion, we will closely follow the latter reference.

We begin by introducing a new notion concerning undirected graphs.

DEFINITION 3.23. Let $G = (V(G), E(G))$ be an undirected graph with vertices $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $X, Y, Z \subseteq V(G)$ be sets of vertices. The set Z is said to separate the set X from the set Y in G , denoted as $\langle X | Z | Y \rangle_G$, if any path from a vertex from X to a vertex from Y involves at least one vertex from Z .

Building on the notion of separation introduced above, we define several types of relationships between probability distributions and undirected graphs.

DEFINITION 3.24. Let G be an undirected graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the Boolean algebra of propositions generated by $\{v_i | V_i \in V(G)\}$. Furthermore, let Pr be a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$. For $X, Y, Z \subseteq V(G)$, let $I_{Pr}(X, Z, Y)$ denote that X is conditionally independent of Y given Z , as defined in Definition 3.18.

- (1) The graph G is called a *dependency map*, or *D-map* for short, of Pr if for all $X, Y, Z \subseteq V(G)$ we have: if $I_{Pr}(X, Z, Y)$ then $\langle X | Z | Y \rangle_G$.
- (2) G is called an *independency map*, or *I-map* for short, of Pr if for all $X, Y, Z \subseteq V(G)$ we have: if $\langle X | Z | Y \rangle_G$ then $I_{Pr}(X, Z, Y)$.
- (3) G is called a *perfect map* of Pr if G is both a *dependency map* and an *independency map* of Pr .

Note that vertices that are adjacent in a D-map of a joint probability distribution Pr are guaranteed to be dependent in Pr (then viewed as statistical variables); the D-map, however, may display a pair of dependent variables as a pair of non-adjacent, that is, separated vertices. On the other hand, vertices found to be non-adjacent in an I-map of Pr correspond to independent

variables; those shown to be adjacent, however, need not necessarily be dependent. A perfect map of Pr faithfully displays all dependencies and independencies embodied in Pr . Not every probability distribution has a perfect map; for some examples see [PEAR88]. It will be evident, however, that every probability distribution has a D-map (an edgeless graph) as well as an I-map (a complete graph).

From now on we will restrict the discussion to I-maps. The following definition introduces the notion of a minimal I-map of a joint probability distribution.

DEFINITION 3.25. *Let G be an undirected graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the Boolean algebra of propositions generated by $\{v_i \mid V_i \in V(G)\}$. Furthermore, let Pr be a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$. The graph G is called a minimal I-map of Pr if G is an I-map of Pr and no proper subgraph of G is.*

Note that a minimal I-map of a joint probability distribution Pr again need not portray all independencies embodied in Pr .

We will shortly see that a joint probability distribution can be represented in terms of functions which are local to the cliques of any one of its I-maps. For this purpose, we introduce the notion of factorization.

DEFINITION 3.26. *Let G be an undirected graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, be the clique set of G . Furthermore, let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$. Let Pr be a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$. Pr is said to factorize according to G if there exist non-negative functions $g_i: \{c_{V(Cl_i)}\} \rightarrow [0, 1]$ such that Pr is defined by*

$$Pr(C_{V(G)}) = \beta \cdot \prod_{i=1, \dots, m} g_i(C_{V(Cl_i)})$$

where β is a normalization factor.

The functions g_i mentioned in the foregoing definition are sometimes called *compatibility functions*, [PEAR88], or *factor potentials*, [LAUR88b]. These compatibility functions are arbitrary in the sense that they are not necessarily related to marginal distributions obtained from the joint probability distribution.

PROPOSITION 3.27. *Let G be an undirected graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Furthermore, let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$. Let Pr be a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$. Then, Pr factorizes according to G if and only if G is an I-map of Pr .*

We refer the reader to [PEAR88] for a proof of the proposition which is originally due to J.H. Hammersley and P. Clifford; in the sequel, we will only use a special case of the proposition.

We have mentioned above that the compatibility functions in a factorization of a joint probability distribution Pr according to an arbitrary I-map in general are not related to marginal distributions derived from Pr . However, if Pr is factorized according to an I-map which is a so-called decomposable graph, then the resulting compatibility functions do have this property. Before we state this more formally in Proposition 3.38, we introduce some new notions concerning graphs.

DEFINITION 3.28. *An undirected graph G is decomposable if all its elementary cycles of length $k \geq 4$ possess a chord.*

Decomposable graphs are also called *triangulated graphs*, [BERG73, LAUR88a], or *chordal graphs*, [PEAR88]; the term decomposable has been adopted from [LAUR84].

DEFINITION 3.29. *Let $G = (V(G), E(G))$ be an undirected graph of order n , $n \geq 1$. Let $\iota: V(G) \leftrightarrow \{1, \dots, n\}$ denote a total ordering of the vertices of G . Now, let the elements of $V(G)$ be numbered V_1, \dots, V_n according to ι (that is, we have $\iota(V_i) = i$). The ordering ι is called a perfect ordering of $V(G)$ if for each $i = 1, \dots, n$, the full subgraph of G induced by the set of vertices $v(V_i) \cap \{V_1, \dots, V_{i-1}\}$ is complete.*

An undirected graph may allow more than one perfect ordering. The notion of a perfect ordering and its definition have been taken from [LAUR88a]. R.E. Tarjan and M. Yannakakis define the notion of a *zero fill-in numbering* and show in a lemma that a total ordering is a zero fill-in numbering if and only if it has the property we have used for a definition, [TARJ84].

The next lemma is of major importance; it has been proven in [TARJ84].

LEMMA 3.30. *Let G be an undirected graph. G is decomposable if and only if it permits a perfect ordering of its vertices.*

The following provides an algorithm for computing a total ordering of the vertices of an arbitrary undirected graph.

ALGORITHM 3.31. *Let $G = (V(G), E(G))$ be an undirected graph of order n , $n \geq 1$. The maximum cardinality search algorithm for computing a total ordering ι of $V(G)$ is the following:*

1. *Assign the number 1 to an arbitrary vertex.*
2. *Number the remaining vertices from 2 to n in increasing order such that the next number is assigned to a vertex having a largest set of previously numbered neighbours.*

Note that in Algorithm 3.31 in each step the next vertex to be numbered need not be unique. Furthermore, the vertex that is numbered need not be a neighbour of the last numbered one.

Tarjan and Yannakakis have proven that when applied to a decomposable graph maximum cardinality search renders a perfect ordering of its vertices:

LEMMA 3.32. *Let G be a decomposable graph. Any ordering ι of the vertices of G obtained from maximum cardinality search is perfect.*

We now define an ordering of the cliques of a decomposable graph.

DEFINITION 3.33. *Let $G = (V(G), E(G))$ be a decomposable graph of order n , $n \geq 1$. Let ι be a perfect ordering of $V(G)$ obtained from maximum cardinality search. Let $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, be the clique set of G . We define the ordering $\hat{\iota}: Cl(G) \leftrightarrow \{1, \dots, m\}$ by $\hat{\iota}(Cl_i) < \hat{\iota}(Cl_j)$ if $\max\{\iota(V_k) \mid V_k \in V(Cl_i)\} < \max\{\iota(V_k) \mid V_k \in V(Cl_j)\}$, for each pair of cliques $Cl_i, Cl_j \in Cl(G)$.*

Note that in the ordering $\hat{\iota}$ introduced above, the cliques of a decomposable graph G are numbered in the order of their highest numbered vertex according to ι . It will be evident that $\hat{\iota}$ is uniquely determined by the ordering ι .

LEMMA 3.34. *Let $G = (V(G), E(G))$ be a decomposable graph. Let ι be a perfect ordering of $V(G)$ obtained from maximum cardinality search. Let $Cl(G)$ be the clique set of G . Let $\hat{\iota}$ be the ordering of $Cl(G)$ obtained from ι as defined above. Then, $\hat{\iota}$ is a total ordering.*

PROOF. The lemma follows from the properties of the ordering ι . ■

The following lemma states an important property of an ordering $\hat{\iota}$ of the cliques of a decomposable graph as defined above; in [LAUR84, TARJ84] further details are provided. The lemma is known as the *running intersection property*.

LEMMA 3.35. *Let G be a decomposable graph. Let $Cl(G)$ be the set of cliques of G numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to an ordering $\hat{\iota}$ as defined in Definition 3.33. Then, $\hat{\iota}$ has the following property: for each $i = 2, \dots, m$, there exists a $j < i$ such that $V(Cl_j) \supset V(Cl_i) \cap (V(Cl_1) \cup \dots \cup V(Cl_{i-1}))$.*

The previous lemma states, in other words, that the vertices a clique has in common with the lower numbered cliques are all contained in one such clique.

DEFINITION 3.36. *Let G be a decomposable graph. Let $Cl(G)$ be the set of cliques of G numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to an ordering $\hat{\iota}$ having the running intersection property. For each $i = 1, \dots, m$, we define $S_i = V(Cl_i) \cap (V(Cl_1) \cup \dots \cup V(Cl_{i-1}))$ with $S_1 = \emptyset$; furthermore, we define $R_i = V(Cl_i) \setminus S_i$. S_i is called the separator of clique Cl_i ; R_i is called its residue.*

We now turn our attention to the factorization of a joint probability distribution according to a decomposable I-map.

DEFINITION 3.37. *A joint probability distribution Pr is called decomposable if it has a minimal I-map that is decomposable. Pr is said to be decomposable relative to an undirected graph G if G is an I-map of Pr and G is decomposable.*

Note that for Pr to be decomposable relative to a graph G , it is not necessary that G is a minimal I-map of Pr .

From the property stated in Proposition 3.27 we conclude that a joint probability distribution Pr which is decomposable relative to a (decomposable) graph G factorizes according to G . In fact, the additional information that G is decomposable allows us to prove the following more specific result.

PROPOSITION 3.38. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $Cl(G)$ be the set of cliques of G numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to an ordering \hat{i} having the running intersection property; for each clique $Cl_i \in Cl(G)$, $i = 1, \dots, m$, let its separator S_i be defined according to Definition 3.36. Let $\mathcal{B}(v_1, \dots, v_n)$ be the Boolean algebra of propositions generated by $\{v_i | V_i \in V(G)\}$; for each $Cl_i \in Cl(G)$, $i = 1, \dots, m$, let $\mathcal{B}(Cl_i) \subseteq \mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_j | V_j \in V(Cl_i)\}$. Furthermore, let Pr be a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$; for $i = 1, \dots, m$, let μ_{Cl_i} be the marginal distribution on $\mathcal{B}(Cl_i)$, derived from Pr . Then, Pr is decomposable relative to G if and only if Pr is defined by*

$$Pr(C_{V(G)}) = \prod_{i=1, \dots, m} \frac{\mu_{Cl_i}(C_{V(Cl_i)})}{\mu_{Cl_i}(C_{S_i})}$$

PROOF.

\Rightarrow We assume that Pr is decomposable relative to G ; we have that the decomposable graph G is an I-map of Pr . So, for all subsets $X, Y, Z \subseteq V(G)$ of vertices such that $\langle X | Z | Y \rangle_G$ we have $I_{Pr}(X, Z, Y)$. Now observe that since \hat{i} has the running intersection property, we have that $\langle V(Cl_i) | S_i | V(Cl_1) \cup \dots \cup V(Cl_{i-1}) \rangle_G$, for $i = 1, \dots, m$; it therefore follows that $Pr(C_{V(Cl_i)} | C_{V(Cl_1) \cup \dots \cup V(Cl_{i-1})}) = Pr(C_{V(Cl_i)} | C_{S_i})$. So,

$$\begin{aligned} Pr(C_{V(G)}) &= \prod_{i=1, \dots, m} Pr(C_{V(Cl_i)} | C_{V(Cl_1) \cup \dots \cup V(Cl_{i-1})}) = \\ &= \prod_{i=1, \dots, m} Pr(C_{V(Cl_i)} | C_{S_i}) = \\ &= \prod_{i=1, \dots, m} \frac{\mu_{Cl_i}(C_{V(Cl_i)})}{\mu_{Cl_i}(C_{S_i})} \end{aligned}$$

⇐ We assume that the joint probability distribution Pr is defined by

$$Pr(C_{V(G)}) = \prod_{i=1, \dots, m} \frac{\mu_{C_{I_i}}(C_{V(C_{I_i})})}{\mu_{C_{I_i}}(C_{S_i})}$$

We have to prove that the decomposable graph G is an I-map of Pr , that is, we have to show that for all $X, Y, Z \subseteq V(G)$ such that $\langle X | Z | Y \rangle_G$ we have $I_{Pr}(X, Z, Y)$. It suffices to show that $Pr(C_{V(C_{I_i})} | C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{i-1}})}) = Pr(C_{V(C_{I_i})} | C_{S_i})$ for $i = 1, \dots, m$. We first prove the statement for $i = m$:

$$\begin{aligned} Pr(C_{V(C_{I_m})} | C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})}) &= \\ &= \frac{Pr(C_{V(G)})}{Pr(C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})})} = \\ &= \frac{Pr(C_{V(G)})}{\sum_{c_{R_m}} Pr(C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})} \wedge c_{R_m})} = \\ &= \frac{\mu_{C_{I_m}}(C_{V(C_{I_m})})}{\sum_{c_{R_m}} \mu_{C_{I_m}}(C_{S_m} \wedge c_{R_m})} = \frac{\mu_{C_{I_m}}(C_{V(C_{I_m})})}{\mu_{C_{I_m}}(C_{S_m})} = Pr(C_{V(C_{I_m})} | C_{S_m}) \end{aligned}$$

Now note that from

$$\begin{aligned} Pr(C_{V(G)}) &= \\ &= Pr(C_{V(C_{I_m})} | C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})}) \cdot Pr(C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})}) = \\ &= \frac{\mu_{C_{I_m}}(C_{V(C_{I_m})})}{\mu_{C_{I_m}}(C_{S_m})} \cdot Pr(C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})}) = \\ &= \prod_{i=1, \dots, m} \frac{\mu_{C_{I_i}}(C_{V(C_{I_i})})}{\mu_{C_{I_i}}(C_{S_i})} \end{aligned}$$

we have that

$$Pr(C_{V(C_{I_i}) \cup \dots \cup V(C_{I_{m-1}})}) = \prod_{i=1, \dots, m-1} \frac{\mu_{C_{I_i}}(C_{V(C_{I_i})})}{\mu_{C_{I_i}}(C_{S_i})}$$

We can now simply repeat the argument for $i = m - 1, \dots, 1$.

■

3.4.2. The Transformation Scheme

In our introduction, we have mentioned that Lauritzen and Spiegelhalter propose transforming the originally assessed belief network into a new representation of the problem domain called a decomposable belief network. Before proceeding with a discussion of the proposed transformation we define the notion of such a decomposable belief network.

DEFINITION 3.39. *A decomposable belief network is a tuple $B = (G, M)$ such that*

- (1) $G = (V(G), E(G))$ is a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, and
- (2) $M = \{\mu_{Cl_i} \mid Cl_i \in Cl(G)\}$ is a set of marginal distributions μ_{Cl_i} on $\mathcal{B}(Cl_i)$ where $\mathcal{B}(Cl_i)$ is the Boolean algebra of propositions generated by $\{v_j \mid V_j \in V(Cl_i)\}$, $i = 1, \dots, m$, such that for each pair of cliques $Cl_i, Cl_j \in Cl(G)$ with $V(Cl_i) \cap V(Cl_j) \neq \emptyset$ we have that $\mu_{Cl_i}(C_{V(Cl_i) \cap V(Cl_j)}) = \mu_{Cl_j}(C_{V(Cl_i) \cap V(Cl_j)})$.

The statement in the following lemma can readily be verified using Proposition 3.38.

LEMMA 3.40. *Let $B = (G, M)$ be a decomposable belief network as defined above. Let $\mathcal{B}(v_1, \dots, v_n)$, $n \geq 1$, be the Boolean algebra of propositions generated by $\{v_i \mid V_i \in V(G)\}$. Then, M defines a joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_n)$ such that Pr is decomposable relative to G .*

The transformation of the originally assessed belief network into a decomposable belief network comprises several steps. These transformation steps are shown in Figure 3.2; the graphical representation of the belief network is transformed into a decomposable graph, and from the probabilistic part of the network a new representation of the joint probability distribution in terms of marginal distributions associated with the cliques of the decomposable graph is obtained. The overall transformation scheme essentially comprises two steps:

- (1) Transform the initial belief network into a so-called moral belief network which consists of a so-called moral graph and an associated representation of the joint probability distribution in terms of so-called evidence potentials.
- (2) Transform the moral belief network into a decomposable belief network which consists of a decomposable graph and an associated representation of the joint probability distribution in terms of marginal distributions.

We discuss these transformation steps in some detail.

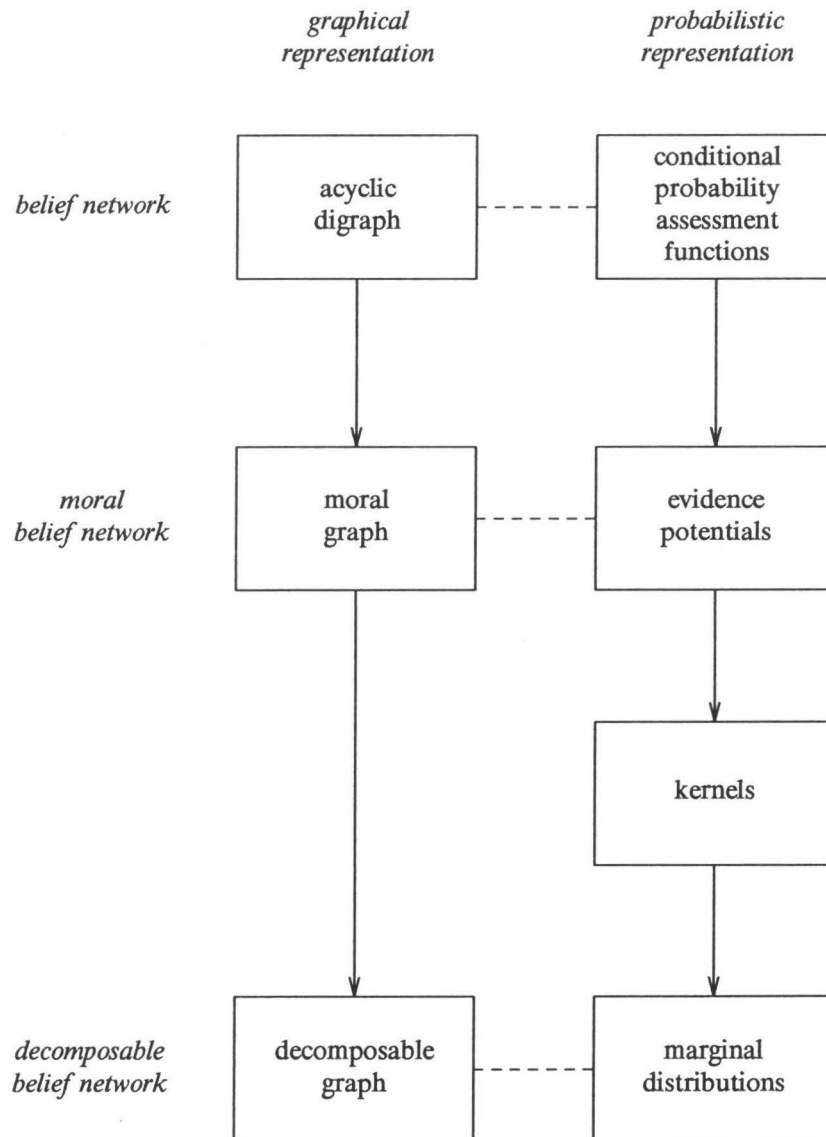


FIGURE 3.2. Transformation of the original belief network.

The Moral Belief Network.

In [SPIE86b], D.J. Spiegelhalter has pointed out that

“the class of recursive models and the class of graphical models intersect in the class of decomposable models, and a recursive model is a member of this intersection provided it does not have two non-adjacent vertices both preceding the same vertex”

([SPIE86b], p. 51; the statement has been proven in [WERM83])

thus providing a motivation and means for the construction of the graphical part of the moral belief network.

Consider a belief network $B = (G, \Gamma)$ as defined in Section 3.2. Informally speaking, the *moral graph* G_M of the acyclic digraph G is obtained by first adding arcs to G such that no vertex in $V(G)$ has non-adjacent predecessors¹, and subsequently dropping the directions of the arcs. The moral graph of an acyclic digraph is defined more formally in the following definition.

DEFINITION 3.41. *Let $G = (V(G), A(G))$ be an acyclic directed graph with vertices $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let H be the (simple) digraph $H = (V(G), A(H))$ such that $A(H) = A(G) \cup \{(V_i, V_j) \mid \text{there is an index } k \text{ such that } V_i, V_j \in \pi(V_k), i < j, \text{ and } V_i \text{ and } V_j \text{ are not adjacent in } G\}$. The moral graph G_M of G is defined as the underlying graph of H .*

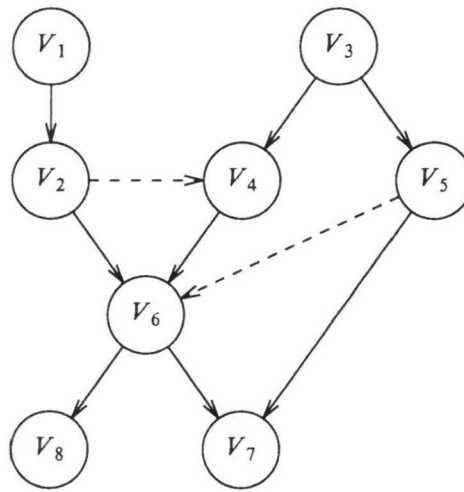
Note that the process of adding arcs between non-adjacent predecessors as described in the previous definition is only performed once: it is not repeated recursively. The construction of the moral graph for our running example is demonstrated in the following example.

EXAMPLE 3.42. Consider the acyclic digraph G from Figure 3.1 once more. Upon successively examining the vertices V_1 up to V_8 inclusive we find that the predecessors of vertex V_6 (that is, the vertices V_2 and V_4) are not adjacent, and that the same holds for the predecessors of vertex V_7 . We therefore add the arcs (V_2, V_4) and (V_5, V_6) to G . Note that the directions of these arcs are irrelevant since we will drop all directions subsequently. The construction of the moral graph G_M of G is shown in Figure 3.3. ■

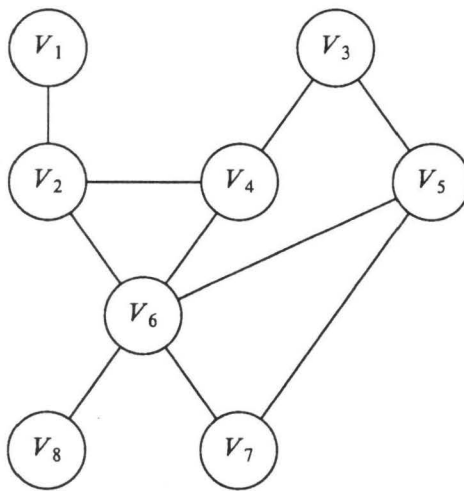
The following lemma will be evident.

LEMMA 3.43. *Let G be an acyclic digraph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. For each $V_i \in V(G)$, let V_i^π be defined as $V_i^\pi = \{V_i\} \cup \pi_G(V_i)$. Furthermore, let G_M be the moral graph of G as defined above. Then for each $V_i \in V(G)$, the full subgraph of G_M induced by V_i^π is complete.*

1. The phrase *moral graph* comes forth from the observation that in a moral graph the parents of a common child are married (that is, there is a direct relationship between them).



(a)



(b)

FIGURE 3.3. Construction of the moral graph G_M .

For the qualitative part of the original belief network we now have obtained an undirected representation. This moral graph again demonstrates certain independency relationships between the statistical variables. The following example shows however that some of the initially assessed independency relationships may no longer be visible explicitly in the moral graph. The moral graph therefore is an I-map but not necessarily a perfect map of the originally assessed joint probability distribution. The arcs that were added to the original graph should be taken as a kind of 'dummy' relationships.

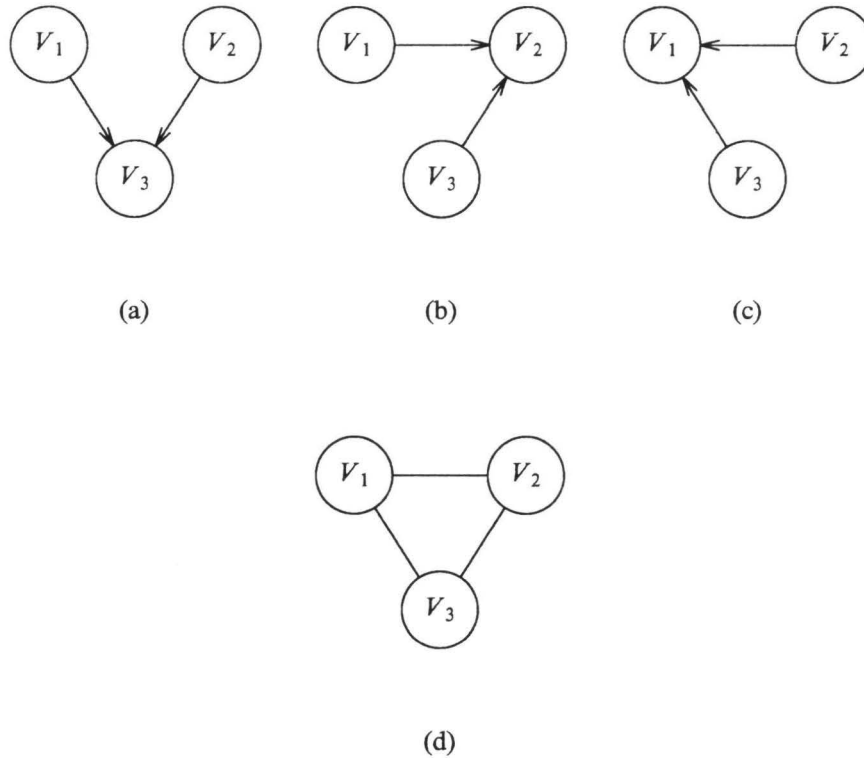


FIGURE 3.4. Three digraphs having the same moral graph.

EXAMPLE 3.44. Consider the graphs shown in Figure 3.4(a) - (d). The digraph (a) for example represents independency of the variables V_1 and V_2 ; this independency however is no longer represented explicitly in the corresponding moral graph (d). The three digraphs (a) - (c) reflect different probabilistic independency relationships between the variables V_1 , V_2 and V_3 . All three digraphs nevertheless have the same moral graph (d). ■

With the obtained moral graph we associate a new ‘undirected’ representation of the initially specified joint probability distribution, again in terms of local factors. This new representation is based on the notion of an *evidence potential*, a real-valued non-negative function of which the values only depend on configurations of small sets of vertices. An evidence potential may be viewed as the proportional contribution of the indicated set of variables to the original joint probability distribution.

Definition 3.45 introduces these evidence potentials and the notion of a moral belief network.

DEFINITION 3.45. Let $B = (G, \Gamma)$ be a belief network as defined in Definition 3.18, where G is an acyclic digraph with vertices $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. For each $V_i \in V(G)$, let V_i^π be defined as $V_i^\pi = \{V_i\} \cup \pi_G(V_i)$. The moral belief network B_M derived from B is the tuple $B_M = (G_M, \Psi)$ where

- (1) G_M is the moral graph of G as defined in Definition 3.41, and
- (2) $\Psi = \{\psi_{V_i^\pi} \mid V_i \in V(G)\}$ is the set of real-valued non-negative functions $\psi_{V_i^\pi}: \{c_{V_i^\pi}\} \rightarrow [0, 1]$, called evidence potentials, such that each $\psi_{V_i^\pi}$ is defined by $\psi_{V_i^\pi}(C_{V_i^\pi}) = \gamma_{V_i}(V_i \mid C_{\pi_G(V_i)})$, $\gamma_{V_i} \in \Gamma$, $C_{V_i^\pi} = V_i \wedge C_{\pi_G(V_i)}$, $i = 1, \dots, n$.

EXAMPLE 3.46. Consider the acyclic digraph G from Figure 3.1 once more and its corresponding moral graph G_M as shown in Figure 3.3(b). With G_M we associate a set of evidence potentials which are obtained from the conditional probability assessment functions associated with G as follows:

$$\begin{aligned}
 \psi_{V_1^\pi}(V_1) &= \gamma_{V_1}(V_1) \\
 \psi_{V_2^\pi}(V_1 \wedge V_2) &= \gamma_{V_2}(V_2 \mid V_1) \\
 \psi_{V_3^\pi}(V_3) &= \gamma_{V_3}(V_3) \\
 \psi_{V_4^\pi}(V_3 \wedge V_4) &= \gamma_{V_4}(V_4 \mid V_3) \\
 \psi_{V_5^\pi}(V_3 \wedge V_5) &= \gamma_{V_5}(V_5 \mid V_3) \\
 \psi_{V_6^\pi}(V_2 \wedge V_4 \wedge V_6) &= \gamma_{V_6}(V_6 \mid V_2 \wedge V_4) \\
 \psi_{V_7^\pi}(V_5 \wedge V_6 \wedge V_7) &= \gamma_{V_7}(V_7 \mid V_5 \wedge V_6) \\
 \psi_{V_8^\pi}(V_6 \wedge V_8) &= \gamma_{V_8}(V_8 \mid V_6)
 \end{aligned}$$

■

It will be evident that the set of evidence potentials constitutes just another representation of the original joint probability distribution defined by the initial assessment functions. This property is stated more formally in the following proposition.

PROPOSITION 3.47. Let $B = (G, \Gamma)$ be a belief network defined according to Definition 3.18, where G is an acyclic digraph with vertices $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. For each $V_i \in V(G)$, let $V_i^\pi = \{V_i\} \cup \pi_G(V_i)$. Furthermore, let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$, and let Pr be the joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$ defined by Γ as in Proposition 3.21. Now, let $B_M = (G_M, \Psi)$ be the moral belief network derived from B defined as above. Then, we have that

$$Pr(C_{V(G)}) = \prod_{V_i \in V(G)} \psi_{V_i^*}(C_{V_i^*})$$

PROOF. The property follows immediately from Definition 3.45 and Proposition 3.21. ■

From Proposition 3.47 we have that although some of the originally assessed independency relationships are no longer explicitly represented in the moral graph, they still are represented in the joint probability distribution. Note that the representation of the joint probability distribution in terms of evidence potentials again is a local representation of uncertainty.

The Decomposable Belief Network.

We recall that the transformation of the initially assessed belief network into a decomposable belief network comprises two steps, the first of which we have discussed just now. We proceed with a discussion of the transformation of the moral belief network resulting from the first transformation step into a corresponding decomposable belief network. We first consider the transformation of the obtained moral graph into a decomposable graph.

The moral graph can be made decomposable by *filling-in*, that is, (once more) by adding certain ‘dummy’ edges. Lauritzen and Spiegelhalter use an efficient algorithm by Tarjan and Yannakakis, [TARJ84], for doing so. The algorithm is known as the fill-in algorithm for obtaining a decomposable graph from an arbitrary undirected one.

ALGORITHM 3.48. Let G be an undirected graph of order n , $n \geq 1$. The fill-in algorithm is the following:

1. Compute a total ordering ι of the vertices of G using maximum cardinality search.
2. From $i = n$ to 1, for each vertex numbered i add edges between any non-adjacent neighbours of i that are assigned a lower number than i in ι .

The set of edges added to G is called the fill-in.

If by applying Algorithm 3.48 no edges are added to an undirected graph G , then G was already decomposable; the phrase ‘zero fill-in numbering’ used by Tarjan and Yannakakis instead of the phrase ‘perfect ordering’ emerges from this observation. Otherwise, the new graph obtained from applying the

algorithm is decomposable. This property is stated more formally in the following lemma. For a proof of the lemma, the reader is referred once more to [TARJ84].

LEMMA 3.49. *Let G be an undirected graph. Let H be an undirected graph obtained from G by using the fill-in algorithm shown above. Then, H is decomposable.*

The following example shows the application of the fill-in algorithm to the moral graph of our running example. Note that the graph which results from applying the algorithm need not be unique.

EXAMPLE 3.50. Consider the moral graph G_M from Figure 3.3(b) once more. We use Algorithm 3.48 to obtain from G_M a decomposable graph G_D . Using maximum cardinality search the vertices of G_M may be numbered as shown in Figure 3.5(a). Examining the vertices from 8 to 1 in decreasing order we find that the vertex numbered 6 has two non-adjacent neighbours that are assigned a lower number than 6 in the ordering: the full subgraph generated by $\{V_4, V_5\} \cap \{V_1, V_2, V_4, V_5, V_6\} = \{V_4, V_5\}$ is not complete. Therefore the edge (V_4, V_5) is added to G_M , yielding the decomposable graph G_D shown in Figure 3.6(b). Note that the alternative addition of (V_3, V_6) would have yielded a decomposable graph as well. We now number the six cliques of G_D in the order of their highest numbered vertex as prescribed in Definition 3.33. Let Cl_i be the clique assigned number i . Then, we have obtained the following ordering \hat{i} (identifying a clique with its vertex set):

$$Cl_1 = \{V_1, V_2\}$$

$$Cl_2 = \{V_2, V_4, V_6\}$$

$$Cl_3 = \{V_4, V_5, V_6\}$$

$$Cl_4 = \{V_3, V_4, V_5\}$$

$$Cl_5 = \{V_5, V_6, V_7\}$$

$$Cl_6 = \{V_6, V_8\}$$

■

In [LAUR88a], Lauritzen and Spiegelhalter point out that the fill-in should be computed very carefully, since the maximal order of the cliques of the decomposable graph resulting from the fill-in determines the computational complexity of their method for evidence propagation. It should be noted that the problem of computing a fill-in containing a minimum number of edges is NP-complete, [YANN81].

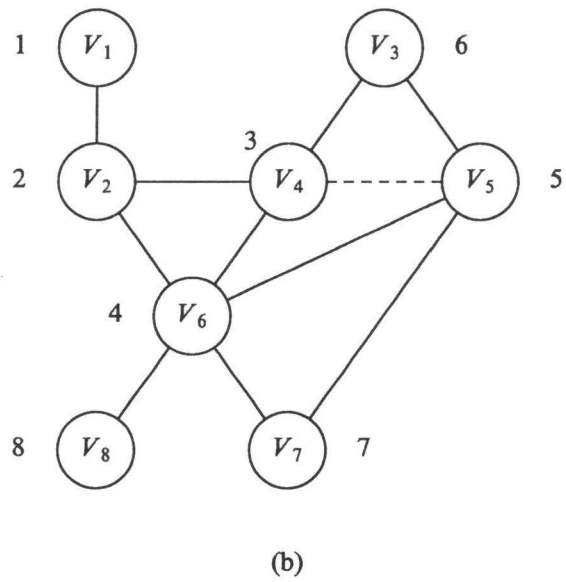
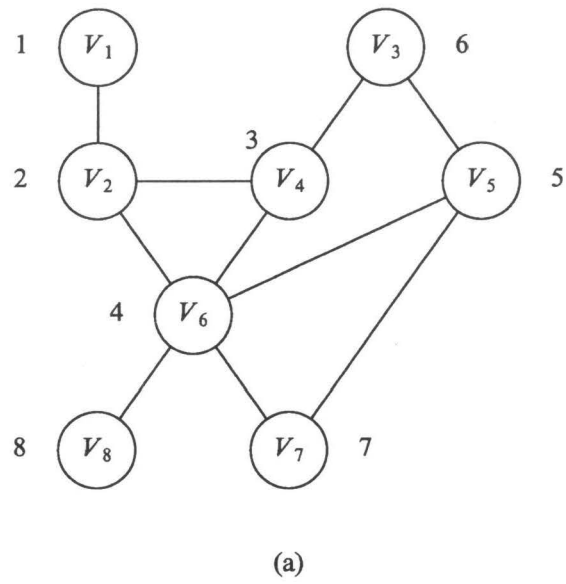


FIGURE 3.5. Construction of the decomposable graph G_D .

The graphical part of the original belief network has now been transformed into a decomposable graph. Recall from Section 3.4.1 that a joint probability distribution which is decomposable relative to a (decomposable) graph G can be expressed in terms of marginal distributions associated with the cliques of G . Lauritzen and Spiegelhalter propose transforming the representation of the joint probability distribution in terms of evidence potentials into such a representation in terms of clique marginals.

We recall that associated with the moral graph G_M we departed from for constructing the decomposable belief network, we had a representation Ψ of the joint probability distribution in terms of evidence potentials ψ_V , $V \subseteq V(G_M)$. We now are interested in a representation of the joint probability distribution in terms of marginal distributions associated with the cliques of a decomposable graph G_D obtained from G_M . Since the evidence potentials ψ_V not necessarily are defined on cliques or clique-intersections, they do not give rise to a representation of the joint probability distribution in clique marginals in a straightforward manner. In Lemma 3.51 we introduce an intermediate representation of the joint probability distribution in terms of so-called kernels; Lemma 3.52 gives the sought-for representation in terms of clique marginals.

LEMMA 3.51. *Let $B_M = (G_M, \Psi)$ be a moral belief network defined according to Definition 3.45 where $V(G_M) = \{V_1, \dots, V_n\}$, $n \geq 1$. Furthermore, let $\mathcal{B}(v_1, \dots, v_n)$ be the Boolean algebra of propositions generated by $\{v_i \mid V_i \in V(G_M)\}$ and let Pr be the joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$ defined by Ψ as in Proposition 3.47. Let G_D be a decomposable graph obtained from G_M by using Algorithm 3.48. Furthermore, let $Cl(G_D)$ be the set of cliques of G_D numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to an ordering i having the running intersection property. For each clique Cl_i , let its separator S_i and its residue R_i be defined as in Definition 3.36. Then, there exists a set $K = \{\kappa_{Cl_i} \mid Cl_i \in Cl(G_D)\}$ of functions $\kappa_{Cl_i}: \{c_{R_i}\} \times \{c_{S_i}\} \rightarrow [0, 1]$ such that*

$$Pr(C_{V(G_M)}) = \prod_{i=1, \dots, m} \kappa_{Cl_i}(C_{R_i} \mid C_{S_i})$$

PROOF. The lemma has been proven by Lauritzen and Spiegelhalter in [LAUR88a]. Since the proof gives a construction of K , we repeat their argument (even in more detail).

Let $D_{m+1} = \{V \mid V \subseteq V(G_M), \psi_V \in \Psi\}$ be the set of (initial) evidence potential domains. We have from Proposition 3.47 that $Pr(C_{V(G_M)}) = \prod_{V \in D_{m+1}} \psi_V(C_V)$ is an evidence potential representation of the

marginal distribution on $\mathcal{B}(v_1, \dots, v_n)$ (or, less formally, on $V(G_M)$). We recursively repeat the following computation for $i = m, \dots, 1$:

Consider clique Cl_i and its residue R_i . We define $V^*(Cl_i) = V(Cl_1) \cup \dots \cup V(Cl_i)$. Now assume that $Pr(C_{V^*(Cl_i)}) = \prod_{V \in D_{i+1}} \psi_V(C_V)$ is an evidence potential representation of the marginal distribution on $V^*(Cl_i)$ (note that the assumption holds for $i = m$). We split the set D_{i+1} into two disjoint subsets: the set $D_i^+ = \{V \mid V \in D_{i+1}, V \cap R_i \neq \emptyset\}$ consisting of those evidence potential domains that contain variables from R_i and the set $D_i^- = D_{i+1} \setminus D_i^+$ consisting of those domains that do not. We have

$$\begin{aligned} Pr(C_{V^*(Cl_i) \setminus R_i}) &= \sum_{c_{R_i}} Pr(C_{V^*(Cl_i) \setminus R_i} \wedge c_{R_i}) = \\ &= \prod_{V \in D_i^-} \psi_V(C_V) \cdot \sum_{c_{R_i}} \left[\prod_{V \in D_i^+} \psi_V(C_{V \setminus R_i} \wedge c_{R_i}) \right] \end{aligned}$$

Note that the first equality merely states that $Pr(C_{V^*(Cl_i) \setminus R_i})$ is obtained from $Pr(C_{V^*(Cl_i)})$ by marginalization; the second equality follows from the earlier mentioned assumption $Pr(C_{V^*(Cl_i)}) = \prod_{V \in D_{i+1}} \psi_V(C_V)$.

We now have obtained a representation of the marginal distribution on $V(Cl_1) \cup \dots \cup V(Cl_{i-1})$. We define a new potential domain $\bar{D}_i = \bigcup_{V \in D_i^+} V \setminus R_i$; we furthermore define the function

$$\phi_{\bar{D}_i}(C_{\bar{D}_i}) = \sum_{c_{R_i}} \left[\prod_{V \in D_i^+} \psi_V(C_{V \setminus R_i} \wedge c_{R_i}) \right]$$

It follows that

$$\begin{aligned} Pr(C_{R_i} \mid C_{S_i}) &= Pr(C_{R_i} \mid C_{V(Cl_1) \cup \dots \cup V(Cl_{i-1})}) = \\ &= \frac{Pr(C_{V(Cl_1) \cup \dots \cup V(Cl_i)})}{Pr(C_{V(Cl_1) \cup \dots \cup V(Cl_{i-1})})} = \\ &= \frac{\prod_{V \in D_{i+1}} \psi_V(C_V)}{\prod_{V \in D_i^-} \psi_V(C_V) \cdot \phi_{\bar{D}_i}(C_{\bar{D}_i})} = \prod_{V \in D_i^+} \frac{\psi_V(C_V)}{\phi_{\bar{D}_i}(C_{\bar{D}_i})} \end{aligned}$$

We add the new evidence potential domain \bar{D}_i to D_i^- as an initialization for the following computation step of the recursion. So, we let

$D_i = D_i^- \cup \{\bar{D}_i\}$; note that the elements of D_i do not contain variables from R_i . We furthermore define a new set of evidence potentials $\bar{\psi}_V$ such that

$$\bar{\psi}_{\bar{D}_i}(C_{\bar{D}_i}) = \psi_{\bar{D}_i}(C_{\bar{D}_i}) \cdot \phi_{\bar{D}_i}(C_{\bar{D}_i}) \text{ in case } \bar{D}_i \in D_{i+1}, \text{ or}$$

$$\bar{\psi}_{\bar{D}_i}(C_{\bar{D}_i}) = \phi_{\bar{D}_i}(C_{\bar{D}_i}), \text{ otherwise}$$

and

$$\bar{\psi}_V = \psi_V \text{ for all } V \in D_{i+1}, V \neq \bar{D}_i$$

resulting in an evidence potential representation of the marginal distribution on $V^*(Cl_{i-1}) = V(Cl_1) \cup \dots \cup V(Cl_{i-1})$. Subsequently, we rename $\bar{\psi}_V$ into ψ_V , and we repeat the computation for $i - 1$. Note that we have established the property $Pr(C_{V^*(Cl_{i-1})}) = \prod_{V \in D_i} \psi_V(C_V)$.

Taking $\kappa_{Cl_i}(C_{R_i} | C_{S_i}) = Pr(C_{R_i} | C_{S_i})$, it will be evident that we find $Pr(C_{V(G_M)}) = \prod_{i=1, \dots, m} \kappa_{Cl_i}(C_{R_i} | C_{S_i})$. ■

The functions κ_{Cl_i} introduced in the preceding lemma are called *kernels*. From these kernels, we now obtain a set of clique marginals. Lemma 3.52 provides a means for doing so.

LEMMA 3.52. *Let $B_M = (G_M, \Psi)$ be a moral belief network defined according to Definition 3.45 where $V(G_M) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i | V_i \in V(G_M)\}$ and let Pr be the joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$ defined by Ψ as in Proposition 3.47. Let G_D be a decomposable graph obtained from G_M by using Algorithm 3.48. Furthermore, let $Cl(G_D)$ be the set of cliques of G_D numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to an ordering \hat{i} having the running intersection property. For each Cl_i , let its separator S_i be defined according to Definition 3.36. Let \mathbb{K} be the set of kernels as constructed in the previous lemma. Then, there exists a set $\mathbb{M} = \{\mu_{Cl_i} | Cl_i \in Cl(G_D)\}$ of marginal distributions μ_{Cl_i} such that*

$$Pr(C_{V(G_M)}) = \prod_{i=1, \dots, m} \frac{\mu_{Cl_i}(C_{V(Cl_i)})}{\mu_{Cl_i}(C_{S_i})}$$

PROOF. The lemma follows immediately by taking $\mu_{Cl_i}(C_{V(Cl_i)}) = \kappa_{Cl_i}(C_{R_i} | C_{S_i})$, and for each Cl_i , $i = 2, \dots, m$, recursively, $\mu_{Cl_i}(C_{V(Cl_i)}) = \kappa_{Cl_i}(C_{R_i} | C_{S_i}) \cdot \mu_{Cl_j}(C_{S_i})$ where $j < i$ is chosen such that $V(Cl_j) \supset V(Cl_i) \cap (V(Cl_1) \cup \dots \cup V(Cl_{i-1}))$ and $\mu_{Cl_j}(C_{S_i})$ is obtained by marginalization. ■

EXAMPLE 3.53. Consider the decomposable graph G_D from Figure 3.5(b) once more. With G_D we associate the set of clique marginals

$$\mathbb{M} = \{\mu_{Cl_1}(C_{\{V_1, V_2\}}), \mu_{Cl_2}(C_{\{V_2, V_4, V_6\}}), \mu_{Cl_3}(C_{\{V_4, V_5, V_6\}}), \\ \mu_{Cl_4}(C_{\{V_3, V_4, V_5\}}), \mu_{Cl_5}(C_{\{V_5, V_6, V_7\}}), \mu_{Cl_6}(C_{\{V_6, V_8\}})\}$$

giving rise to the following representation of the joint probability distribution:

$$\begin{aligned} Pr(V_1 \wedge \dots \wedge V_8) = \\ = \mu_{Cl_1}(V_1 \wedge V_2) \cdot \frac{\mu_{Cl_2}(V_2 \wedge V_4 \wedge V_6)}{\mu_{Cl_2}(V_2)} \cdot \frac{\mu_{Cl_3}(V_4 \wedge V_5 \wedge V_6)}{\mu_{Cl_3}(V_4 \wedge V_6)} \cdot \\ \cdot \frac{\mu_{Cl_4}(V_3 \wedge V_4 \wedge V_5)}{\mu_{Cl_4}(V_4 \wedge V_5)} \cdot \frac{\mu_{Cl_5}(V_5 \wedge V_6 \wedge V_7)}{\mu_{Cl_5}(V_5 \wedge V_6)} \cdot \frac{\mu_{Cl_6}(V_6 \wedge V_8)}{\mu_{Cl_6}(V_6)} \end{aligned}$$

■

The transformation of the initially assessed belief network into a decomposable belief network has now been described completely. We have mentioned before that the actual scheme for evidence propagation proposed by Lauritzen and Spiegelhalter operates on this decomposable belief network. We emphasize that for a specific problem domain the transformation needs to be performed only once.

3.4.3. Evidence Propagation in a Decomposable Belief Network

For making probabilistic statements concerning the statistical variables discerned in a problem domain we have to associate with a decomposable belief 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 belief network is concerned, it will be evident that any probability which involves variables occurring in one and the same clique only, can simply be computed locally from the marginal distribution on that clique.

The method for evidence propagation is less straightforward. Suppose that evidence becomes available that a statistical variable V has adopted a certain value, say *true*. For ease of exposition, we assume that the variable V occurs in only one clique of the decomposable graph; our observations, however, can easily be adapted to deal with the more general case in which V occurs in the intersection of two or more cliques. Informally speaking, propagation of this piece of evidence amounts to the following. The vertices and the cliques of the graph of the decomposable belief network are ordered anew as described in Section 3.4.1, this time starting with the instantiated vertex. The ordering of the cliques then is taken as the order in which the evidence is processed through the cliques; for each subsequent clique, the updated marginal

distribution is computed locally. Then, the instantiated vertex is removed from the graph, and the updated marginal distributions are taken as the marginal distributions associated with the cliques of the remaining graph, together once more constituting a decomposable belief network.

In Definition 3.53 we define the set M^v of updated marginal distributions. In Proposition 3.54 we state that M^v indeed defines the updated probability distribution given $V = \text{true}$.

DEFINITION 3.53. Let $B = (G, M)$ be a decomposable belief network. Let $Cl(G)$ be the clique set of G . Now let the evidence $V = \text{true}$ (or $V = \text{false}$, alternatively) be observed for a vertex $V \in V(G)$. Let ι be an ordering of $V(G)$ obtained from maximum cardinality search starting with V (so, we have $\iota(V) = 1$). Let the cliques of G be numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to the ordering $\hat{\iota}$ obtained from ι as in Definition 3.33. For each Cl_i , let its separator S_i be defined according to Definition 3.36. We define the updated marginal distributions $\mu_{Cl_i}^v$, $i = 1, \dots, m$, as follows:

- (1) $\mu_{Cl_1}^v(C_{Cl_1 \setminus \{V\}}) = \mu_{Cl_1}(C_{Cl_1 \setminus \{V\}} \wedge v)$, and
- (2) $\mu_{Cl_i}^v(C_{Cl_i}) = \mu_{Cl_i}(C_{Cl_i}) \cdot \frac{\mu_{Cl_i}^v(C_{S_i})}{\mu_{Cl_i}(C_{S_i})}$, for $i = 2, \dots, m$, recursively.

We use M^v to denote the set of updated marginal distributions $M^v = \{\mu_{Cl_i}^v \mid i = 1, \dots, m\}$.

PROPOSITION 3.54. Let $B = (G, M)$ be a decomposable belief network where $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $Cl(G)$ be the clique set of G . Now, let the evidence $V = \text{true}$ (or $V = \text{false}$, alternatively) be observed for a vertex $V \in V(G)$. Let ι be an ordering of $V(G)$ obtained from maximum cardinality search starting with V . Let the cliques of G be numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to $\hat{\iota}$ obtained from the ordering ι as in Definition 3.33. For each Cl_i , let its separator S_i be defined as in Definition 3.36. Let $\mathcal{B}(v_1, \dots, v_n)$ be the Boolean algebra of propositions generated by $\{v_i \mid V_i \in V(G)\}$ and let Pr be the joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$ defined by M as in Proposition 3.38. Let Pr^v be the updated probability distribution given v . Now, let M^v be as in the previous definition. Then,

$$Pr^v(C_{V(G) \setminus \{V\}}) = \mu_{Cl_1}^v(C_{Cl_1 \setminus \{V\}}) \cdot \prod_{i=2, \dots, m} \frac{\mu_{Cl_i}^v(C_{Cl_i})}{\mu_{Cl_i}^v(C_{S_i})}$$

From Definition 3.53 and Proposition 3.54 we have that propagation of evidence entails only local updating of the joint probability distribution. Note the analogy with the locality of the schemes for propagating evidence in the quasi-probabilistic models discussed in Chapter 2.

The following lemma states that the notion of a decomposable belief network is invariant under evidence propagation.

LEMMA 3.55. *Let $B = (G, M)$ be a decomposable belief network. Let the evidence $V = \text{true}$ (or $V = \text{false}$, alternatively) be observed for a vertex $V \in V(G)$. Let G^v be defined as $G^v = (V(G^v), E(G^v))$ such that $V(G^v) = V(G) \setminus \{V\}$ and $E(G^v) = E(G) \cap (V(G^v) \times V(G^v))$. Furthermore, let M^v be defined as in Definition 3.53. Then, $B^v = (G^v, M^v)$ is a decomposable belief network.*

PROOF. The lemma follows from Proposition 3.54 and the observation that a decomposable graph remains decomposable after removing an arbitrary vertex from it in the way mentioned in the lemma. ■

The method presented by Lauritzen and Spiegelhalter has served as a point of departure for the implementation of a network-based expert system shell called HUGIN at Aalborg University, Denmark, [ANDE89]. HUGIN offers, among other features, a set of tools for constructing a belief network, for subsequently transforming it into a decomposable belief network and for entering and propagating evidence.

Chapter 4

Partially Quantified Belief Networks

In Chapter 1 we have argued that one of the problems in applying probability theory in a model for handling uncertainty in a knowledge-based system is the difficulty of obtaining a joint probability distribution on the problem domain: often only a few probabilities are known or can be estimated by an expert in the field. In such a case, we are confronted with the problem of having to derive statements concerning probabilities of interest from only a partial and often inconsistent specification of a joint probability distribution. We have seen in Chapter 2 that the quasi-probabilistic models developed in the 1970s were able to handle this problem, although not in a mathematically sound way: we have shown that the schemes for combining and propagating evidence employed in these models are incorrect. In fact, even as an approximation technique they are far from convincing.

In contrast, the schemes for evidence propagation employed in the network models are mathematically sound. These models, however, are not capable of dealing with a partial specification of a joint probability distribution nor with an inconsistent one: in the models presented so far the belief network has to be fully and consistently quantified, that is, the initially assessed local (conditional) probabilities have to define a unique joint probability distribution on the statistical variables concerned. Several contributors to the discussion of the paper by S.L. Lauritzen and D.J. Spiegelhalter have called attention to the difficulty of assessing all probabilities required, see for example the contributions by P. Cheeseman and F. Critchley, [LAUR88a]. In the same discussion, D. Dubois and H. Prade furthermore argue that the requirement for a unique joint probability distribution on the statistical variables almost inevitably leads to replacing missing information by strong default

assumptions concerning independency relationships between the variables in order to be able to guarantee uniqueness, with all the unpleasant consequences we encountered before with the quasi-probabilistic models.

In this chapter we address the problem of having only a *partially quantified belief network* at our disposal for reasoning with uncertainty. Informally speaking, a partially quantified belief network is a kind of belief network: it equally consists of a qualitative representation of a problem domain in terms of an acyclic digraph, and a quantification of the arcs of the graph in terms of local probabilities. In a partially quantified belief network, however, the initially given local probabilities do not give rise to a unique joint probability distribution respecting the independency relationships between the statistical variables shown in the graph. We have chosen the phrase *partially quantified belief network* so as to express that only the quantitative part of the representation of a problem domain has been specified partially: we assume that the qualitative part has been fully specified, that is, we assume that all 'directed' independency relationships between the statistical variables are known or have been assessed by an expert in the field.

Now recall from Chapter 3 that for making probabilistic statements concerning the statistical variables discerned in the problem domain, two methods were associated with a belief network: a method for deriving from the network information about probabilities of interest and a method for processing evidence. In order to be able to exploit a partially quantified belief network for reasoning with uncertainty, we have to devise similar methods for this type of belief network. In Section 4.3 we will present a method for computing upper and lower bounds on probabilities of interest from a partially quantified belief network. The general idea of our method is to take the probabilities provided by a domain expert as defining constraints on a yet unknown probability distribution; Section 4.2 provides the basic details of this approach. For exploiting independency relationships between the statistical variables discerned we build on the model by Lauritzen and Spiegelhalter as discussed in the preceding chapter. In Section 4.4 we once more take this model as a starting point for investigating the problem of evidence propagation in a partially quantified belief network. Adhering to the basic idea of the model by Lauritzen and Spiegelhalter, our aim was to arrive at a method for processing evidence in which the graphical part of the network is exploited more or less directly as a computational architecture and which renders the notion of a partially quantified belief network invariant under evidence propagation. Unfortunately we have not been able to devise such a propagation method; in fact, by means of a counterexample we will show that the method for evidence propagation proposed by Lauritzen and Spiegelhalter cannot be extended in a straightforward manner to deal with probability intervals. We will comment on some of the problems we encountered in trying to meet the mentioned requirements.

4.1. PRELIMINARIES: THE THEORY OF LINEAR PROGRAMMING

In this section we review some basic notions from the theory of linear programming. For further information, the reader is referred to [PAPA82,SCHR86]; he may consult [STRA76] for information on aspects from linear algebra and [BRØN83] for more definitions and properties concerning convex polytopes.

DEFINITION 4.1. *Let $x, y \in \mathbb{R}^n$, $n \geq 1$. A convex combination of x and y is a vector $z \in \mathbb{R}^n$ such that $z = \lambda x + (1 - \lambda)y$, $\lambda \in \mathbb{R}$, $0 \leq \lambda \leq 1$.*

A set $S \subseteq \mathbb{R}^n$ is convex if for each $x, y \in S$, all convex combinations of x and y are in S . A convex set is called a (convex) polyhedron if it is the intersection of a finite number of closed halfspaces; if a polyhedron is bounded and nonempty, it is called a (convex) polytope.

For any subset $S \subseteq \mathbb{R}^n$, the convex hull of S , denoted by $\text{hull}(S)$, is the set obtained by recursively taking convex combinations starting with the elements from S . We say that S spans the convex hull $\text{hull}(S)$.

Note that for each set $S \subseteq \mathbb{R}^n$, the convex hull $\text{hull}(S)$ is the smallest convex set containing S . Furthermore, it will be evident that a convex set is a polytope if and only if it is the convex hull of a finite nonempty set $S \subseteq \mathbb{R}^n$.

DEFINITION 4.2. *Let $x, y \in \mathbb{R}^n$, $n \geq 1$. A conical combination of x and y is a vector $z \in \mathbb{R}^n$ such that $z = \lambda x + \mu y$, $\lambda, \mu \geq 0$.*

A set $S \subseteq \mathbb{R}^n$ is a cone if for each $x \in S$ and $\lambda \in \mathbb{R}$, $\lambda \geq 0$, we have $\lambda x \in S$. For a given $x \in S$, the halfline λx , $\lambda \geq 0$, is called the ray spanned by x .

A convex cone is a cone which is convex. A convex cone is called polyhedral if it is the intersection of a finite number of linear halfspaces.

For any subset $S \subseteq \mathbb{R}^n$, the convex cone generated by S , denoted by $\text{cone}(S)$, is the set obtained by recursively taking conical combinations starting with the elements from S .

DEFINITION 4.3. *Let $P \subseteq \mathbb{R}^n$, $n \geq 1$, be a nonempty convex polyhedron and let H be a closed halfspace defined by a hyperplane h . If the intersection $f = P \cap H$ is a subset of h , then f is called a face of P and h is called the supporting hyperplane defining f .*

A face of P of dimension $n - 1$ is called a facet of P . A face of dimension one is called an edge. A face of dimension zero is called a vertex. We use $\text{vert}(P)$ to denote the set of vertices of P .

Informally speaking, a supporting hyperplane defining a facet of a convex polyhedron corresponds to a defining hyperplane of the polyhedron, a vertex is a 'corner' of the polyhedron and an edge is a line segment joining two vertices. A point on a face of a convex polyhedron will be called an *extreme point*; a

point of a convex polyhedron that is not extreme will be called an *interior point*.

It will be evident that a convex polytope has finitely many faces. In the sequel, we will use the following important property.

LEMMA 4.4. *A convex polytope is the convex hull of its vertices.*

In the following definition we define a linear programming problem.

DEFINITION 4.5. *A linear programming problem (or LP-problem for short) in general form is a problem having the following form:*

$$\text{maximize } \sum_{j=1}^n c_j x_j$$

subject to

- (i) $\sum_{j=1}^n a_{ij} x_j \leq b_i, \text{ for } i = 1, \dots, k, k \geq 0,$
- (ii) $\sum_{j=1}^n a_{ij} x_j = b_i, \text{ for } i = k+1, \dots, m, m \geq k,$
- (iii) $x_j \geq 0, \text{ for } j = 1, \dots, n, n \geq 1,$

where the constants a_{ij} constitute the $m \times n$ matrix A , the variables x_j constitute the n -vector \mathbf{x} , the constants b_i constitute the m -vector \mathbf{b} and the constants c_j constitute the vector \mathbf{c} . The linear function to be maximized is called the objective function of the linear programming problem. The equalities and inequalities (i), (ii) and (iii) are called the constraints of the problem; the constraints (iii) are also called the nonnegativity constraints. A system of constraints is called homogeneous if $\mathbf{b} = \mathbf{0}$; otherwise it is called an inhomogeneous system of constraints.

An LP-problem involving inequalities only is said to be in canonical form; an LP-problem involving only equalities and nonnegativity constraints is said to be in standard form.

A vector \mathbf{x} satisfying a system of constraints is called a feasible solution to the system. The set of all feasible solutions to a system of constraints is called its feasible set. A feasible solution that maximizes a given objective function is called an optimal solution; the corresponding value of the objective function is called the optimal value. If a system of constraints has no feasible solutions at all it is called an infeasible system of constraints.

The phrases general form, canonical form and standard form which in the previous definition have been related to LP-problems often are taken to apply to systems of constraints as well. It can easily be shown that the three forms are equivalent, that is, a system of constraints in one form can be transformed into one in another form having the same feasible set.

LEMMA 4.6. *The feasible set of a system of linear constraints is a convex polyhedron. The feasible set of a homogeneous system of linear constraints is a polyhedral cone.*

Definition 4.7 defines the notion of an ϵ -neighbourhood in relation with a system of linear constraints in standard form. Because of the equivalence of the three forms, the definition can be reformulated to apply to systems of constraints in one of the other forms.

DEFINITION 4.7. *Let $Ax = b$, $x \geq 0$, be a system of linear constraints in standard form and let P be its feasible set. For each $y \in P$ and $\epsilon > 0$, the ϵ -neighbourhood of y , denoted by $N_\epsilon(y)$, is defined as the set $N_\epsilon(y) = \{x \mid x \in P \text{ and } \|y - x\| \leq \epsilon\}$, where $\|\cdot\|$ denotes Euclidean distance.*

From the theory of linear programming we have the following theorem: the objective function of an LP-problem having the convex polytope P as the feasible set of its system of constraints assumes its optimal value at a vertex of P . From this theorem it follows that we only have to consider the vertices of P to determine an optimal solution. A well-known computation procedure for solving linear programming problems exploiting this property is the *simplex method* which has been developed by G.B. Dantzig. We do not consider this method in detail; further details can be found in [PAPA82]. The worst-case behaviour of the simplex method is exponential in the size of the LP-problem being solved, which in turn is dependent upon the number of variables the problem involves as well as the number of constraints. Besides this simplex method several other computational methods for solving linear programming problems have been developed. Some of these have been shown to be polynomial in time, such as the *ellipsoid method* presented by L.G. Kachian. This method is discussed in detail in [PAPA82]. Complexity properties of computational methods for solving LP-problems have been formulated in considerable detail by A. Schrijver in [SCHR86].

4.2. PARTIAL SPECIFICATION OF A JOINT PROBABILITY DISTRIBUTION

As has been mentioned before, in this chapter we will deal with the situation in which only a partially quantified belief network is available for reasoning with uncertainty. Recall that such a partially quantified belief network was meant to consist of an acyclic digraph representing independency relationships between the statistical variables discerned, and a partial specification of a joint probability distribution on the variables comprising prior as well as conditional probabilities assessed by a domain expert. In this section we lay the foundation for a method for deriving bounds on probabilities of interest from such a partially quantified belief network. We will concentrate ourselves on the notion of a partial specification of a joint probability distribution and develop a method for calculating probability intervals given such a partial specification. For the moment, we do not take the graphical part of the

network into consideration; in fact, we assume that no independency relationships hold between the statistical variables discerned. In Section 4.3, however, we will present a method for computing bounds on probabilities in which the independencies portrayed in the graphical part of the partially quantified belief network are exploited.

In the following definition the notion of a partial specification of a joint probability distribution is formally defined.

DEFINITION 4.8. *Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. A partial specification of a joint probability distribution on \mathcal{B} is a total function $P: \mathcal{C} \rightarrow [0, 1]$ where $\mathcal{C} \subseteq \mathcal{B}$.*

A partial specification $P: \mathcal{C} \rightarrow [0, 1]$ is consistent if there exists at least one joint probability distribution Pr on \mathcal{B} such that $Pr|_{\mathcal{C}} = P$; otherwise, P is said to be inconsistent. Furthermore, we say that P (uniquely) defines Pr , or alternatively that P is a definition for Pr , if Pr is the only joint probability distribution on \mathcal{B} such that $Pr|_{\mathcal{C}} = P$.

We note that D.V. Lindley et al. use the terms *coherent* and *incoherent* instead of *consistent* and *inconsistent*, [LIND79]. In this section we will often use the incomplete phrase partial specification to denote a partial specification of a joint probability distribution on a given Boolean algebra of propositions as long as ambiguity cannot occur.

The problem of determining the probability of an event given a partial specification of a joint probability distribution has already been investigated as early as halfway the nineteenth century by G. Boole, [BOOL54]. However, Boole's ideas on probability theory have received little attention. In an excellent book providing a thorough exposition of Boole's work on logic and probability in terms of modern algebra, propositional logic and probability theory, T. Hailperin states the following:

"Never clearly understood, and considered anyhow to be wrong, Boole's ideas on probability were simply by-passed by the history of the subject, which developed along other lines."
([HAIL86], p. 215)

In our opinion Boole's ideas have become topical once more in the context of reasoning with uncertainty in artificial intelligence. In fact, our method for deriving mathematically sound statements concerning probabilities of interest from a partial specification of a joint probability distribution is based on Boole's work; we have used Hailperin's book, [HAIL86], as a guide to the work of Boole.

In Section 4.2.1 we will present our method for computing bounds on probabilities of interest from a consistent partial specification of a joint probability distribution. For ease of exposition, we will assume that such a partial specification only comprises prior probabilities; in Section 4.2.2 it will be shown, however, that the method we have developed can deal with conditional probabilities in the same way in which it handles prior ones. In

Section 4.2.3 we will briefly touch upon the problem of having an inconsistent partial specification of a joint probability distribution.

4.2.1. Computing Bounds on Probabilities of Interest

In the following definition, we introduce the notion of a basis for a joint probability distribution. This notion will play an important role in the remainder of this section.

DEFINITION 4.9. Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. A set $\mathcal{C} \subseteq \mathcal{B}$ is called a *basis for a joint probability distribution on \mathcal{B}* if for any consistent partial specification $P: \mathcal{C} \rightarrow [0, 1]$ defined on \mathcal{C} , there exists a joint probability distribution Pr on \mathcal{B} such that P is a definition for Pr .

In Definition 4.10 we introduce a basis that will be shown to have some convenient properties shortly.

DEFINITION 4.10. Let $\mathcal{A} = \{a_1, \dots, a_n\}$, $n \geq 1$, be a set of atomic propositions and let $\mathcal{B}(a_1, \dots, a_n)$ be the Boolean algebra of propositions generated by \mathcal{A} as defined in Definition 3.7. We define the set $\mathcal{B}_0 \subseteq \mathcal{B}(a_1, \dots, a_n)$ such that $\mathcal{B}_0 = \{ \bigwedge_{i=1}^n L_i \mid L_i = a_i \text{ or } L_i = \neg a_i, a_i \in \mathcal{A} \}$.

Note that the set \mathcal{B}_0 has been introduced before in the proof of Proposition 3.12. Recall that it has 2^n elements essentially being the ‘smallest’ ones from $\mathcal{B}(a_1, \dots, a_n)$. It can easily be shown that \mathcal{B}_0 indeed is a basis.

LEMMA 4.11. Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7 and let the set $\mathcal{B}_0 \subseteq \mathcal{B}$ be defined as above. Then, \mathcal{B}_0 is a basis for a joint probability distribution on \mathcal{B} .

The basis $\mathcal{B}_0 \subseteq \mathcal{B}$ will be used frequently throughout the remainder of this chapter. Note that by definition we have that each consistent partial specification $P: \mathcal{B}_0 \rightarrow [0, 1]$ defined on \mathcal{B}_0 , uniquely defines a joint probability distribution Pr on the entire Boolean algebra of propositions \mathcal{B} .

It will be evident that in \mathcal{B} we can identify several different bases. The following lemma for example states another two sets that can easily be shown to be bases for a joint probability distribution on \mathcal{B} .

LEMMA 4.12. Let $\mathcal{A} = \{a_1, \dots, a_n\}$, $n \geq 1$, be a set of atomic propositions, and let $\mathcal{B}(a_1, \dots, a_n)$ be the Boolean algebra of propositions generated by \mathcal{A} . Then, the set $\{ \bigwedge_{i \in \mathcal{J}} a_i \mid \mathcal{J} \subseteq \{1, \dots, n\}, a_i \in \mathcal{A} \}$ is a basis for a joint probability distribution on $\mathcal{B}(a_1, \dots, a_n)$, and so is $\{ \bigvee_{i \in \mathcal{J}} a_i \mid \mathcal{J} \subseteq \{1, \dots, n\}, a_i \in \mathcal{A} \}$.

PROOF. We only prove the lemma for $\{ \bigwedge_{i \in \mathcal{J}} a_i \mid \mathcal{J} \subseteq \{1, \dots, n\}, a_i \in \mathcal{A} \}$. The proof for the set $\{ \bigvee_{i \in \mathcal{J}} a_i \mid \mathcal{J} \subseteq \{1, \dots, n\}, a_i \in \mathcal{A} \}$ follows by symmetry.

Let $\mathcal{C} = \{ \bigwedge_{i \in \mathcal{J}} a_i \mid \mathcal{J} \subseteq \{1, \dots, n\}, a_i \in \mathcal{A} \}$ and let $P: \mathcal{C} \rightarrow [0, 1]$ be a consistent partial specification of a joint probability distribution on $\mathcal{B}(a_1, \dots, a_n)$. Using P , we now construct a (total) joint probability distribution Pr on $\mathcal{B}(a_1, \dots, a_n)$ such that $Pr|_{\mathcal{C}} = P$. By definition we have $Pr(c) = P(c)$ for all $c \in \mathcal{C}$, that is, the probabilities $Pr(c)$ coincide with the initially specified function values $P(c)$. First, the probabilities of conjunctions comprising negated elements of \mathcal{A} are determined uniquely by recursively applying the rule

$$Pr(c \wedge \neg a) = Pr(c) - Pr(c \wedge a)$$

for all $c \in \mathcal{C}$, $a \in \mathcal{A}$. Note that this rule has been derived from Definition 3.11. The function values $Pr(x)$ for all other elements $x \in \mathcal{B} \setminus \mathcal{C}$ are now determined uniquely by recursively using the following properties from Lemma 3.13:

- (1) for all $x_1, x_2 \in \mathcal{B}$, $Pr(x_1 \vee x_2) + Pr(x_1 \wedge x_2) = Pr(x_1) + Pr(x_2)$,
and
- (2) for all $x \in \mathcal{B}$, $Pr(x) + Pr(\neg x) = 1$.

Note that the joint probability distribution Pr is computed using P only and therefore is a unique extension of P . Since P is an arbitrary consistent partial specification defined on \mathcal{C} , we have that \mathcal{C} is a basis for a joint probability distribution on \mathcal{B} . ■

The following lemma can easily be proven.

LEMMA 4.13. *Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$, as defined in Definition 3.7. Then, a basis for a joint probability distribution on \mathcal{B} has at least $2^n - 1$ elements.*

Note that it does not follow from Lemma 4.13 that when less than $2^n - 1$ probabilities have been specified initially, they cannot define a joint probability distribution on a Boolean algebra of propositions \mathcal{B} with n free generators uniquely: it may well be that a consistent partial specification P defined on a subset $\mathcal{C} \subseteq \mathcal{B}$ with $|\mathcal{C}| < 2^n - 1$ is a definition for a joint probability distribution on \mathcal{B} .

EXAMPLE 4.14. Let $\mathcal{A} = \{a_1, a_2, a_3\}$ be a set of atomic propositions and let $\mathcal{B}(a_1, a_2, a_3)$ be the Boolean algebra of propositions generated by \mathcal{A} . Now, let $\mathcal{C} = \{a_1 \wedge a_2 \wedge a_3\}$. It will be evident that the partial specification P defined on \mathcal{C} by $P(a_1 \wedge a_2 \wedge a_3) = 1$ is a definition for a joint probability distribution on the entire Boolean algebra $\mathcal{B}(a_1, a_2, a_3)$. The set \mathcal{C} however is not a basis for a probability distribution on $\mathcal{B}(a_1, a_2, a_3)$. ■

We introduce the notion of a minimal basis.

DEFINITION 4.15. Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$. Let $\mathcal{C} \subseteq \mathcal{B}$. \mathcal{C} is called a *minimal basis* for a joint probability distribution on \mathcal{B} if \mathcal{C} is a basis as defined in Definition 4.9 and if in addition we have $|\mathcal{C}| = 2^n - 1$.

COROLLARY 4.16. Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7 and let $\mathcal{B}_0 \subseteq \mathcal{B}$ be the basis defined according to Definition 4.10. Then, \mathcal{B}_0 is not a minimal basis.

Note that the basis \mathcal{B}_0 contains just one element too many to be a minimal basis. For, since the Boolean algebra of propositions \mathcal{B} is finite we have for each joint probability distribution Pr on \mathcal{B} that any probability $Pr(b_i)$, $b_i \in \mathcal{B}_0$, can be expressed in terms of the probabilities of all other elements from \mathcal{B}_0 : $Pr(b_i) = 1 - \sum_{j=1, j \neq i}^{2^n} Pr(b_j)$. The deletion of an arbitrary element from \mathcal{B}_0 therefore yields a minimal basis.

The following three lemmas state some general properties concerning the basis \mathcal{B}_0 .

LEMMA 4.17. Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$, as defined in Definition 3.7. Let the basis $\mathcal{B}_0 \subseteq \mathcal{B}$ be defined according to Definition 4.10 and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Then, for any joint probability distribution Pr on \mathcal{B} we have

$$\sum_{i=1}^{2^n} Pr(b_i) = 1$$

PROOF. From our definition of the basis \mathcal{B}_0 we have for any $i, j \in \{1, \dots, 2^n\}$ with $i \neq j$ that $b_i \wedge b_j = \text{false}$. Furthermore, we have that $\bigvee_{i=1}^{2^n} b_i = \text{true}$. The result now follows from the observation $Pr(\text{true}) = 1$ and the additivity of any probability distribution Pr on \mathcal{B} . ■

The probabilities $Pr(b_i)$ of $b_i \in \mathcal{B}_0$, $i = 1, \dots, 2^n$, as mentioned in the previous lemma will be called the *constituent probabilities* of Pr .

LEMMA 4.18. Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$, as defined in Definition 3.7. Let the basis $\mathcal{B}_0 \subseteq \mathcal{B}$ be defined according to Definition 4.10 and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Then, for each $b \in \mathcal{B}$ there exists a unique set of indices $\mathcal{J}_b \subseteq \{1, \dots, 2^n\}$ such that $b = \bigvee_{i \in \mathcal{J}_b} b_i$.

PROOF. Each element $b \in \mathcal{B}$ can be written in disjunctive normal form, that is, b can be represented uniquely as a disjunction of elements of \mathcal{B}_0 using De Morgan's laws and the distributive laws. For further details, the reader is referred to the proof of Proposition 3.12. So, there is a unique set of indices $\mathcal{J}_b \subseteq \{1, \dots, 2^n\}$ such that $b = \bigvee_{i \in \mathcal{J}_b} b_i$. ■

The unique set of indices \mathcal{I}_b for an element $b \in \mathcal{B}$ having the property mentioned in the preceding lemma will be called the *index set* for b .

LEMMA 4.19. *Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$. Let the basis $\mathcal{B}_0 \subseteq \mathcal{B}$ be defined as in the foregoing and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Furthermore, let $b \in \mathcal{B}$ and let \mathcal{I}_b be the index set for b as in Lemma 4.18. Then, for each joint probability distribution Pr on \mathcal{B} we have*

$$Pr(b) = \sum_{i \in \mathcal{I}_b} Pr(b_i)$$

PROOF. From Lemma 4.18 we have that $b = \bigvee_{i \in \mathcal{I}_b} b_i$. From the additivity of Pr and the observation that $b_i \wedge b_j = \text{false}$ for any $i, j \in \{1, \dots, 2^n\}$, $i \neq j$, we derive the property stated in the lemma. ■

We will exploit the set \mathcal{B}_0 and its properties for computing probability intervals for probabilities of interest from an arbitrary partial specification. Suppose that we are given probabilities for a number of arbitrary Boolean combinations of atomic propositions, that is, we consider the case in which we are given a consistent partial specification P of a joint probability distribution on \mathcal{B} , which is defined on an arbitrary subset $\mathcal{C} \subseteq \mathcal{B}$. The problem of finding a joint probability distribution on \mathcal{B} which is an extension of P will now be transformed into an equivalent problem in linear algebra. The general idea is to take the initially given probabilities as defining constraints on a yet unknown joint probability distribution.

Let \mathcal{B} once more be a Boolean algebra of propositions with n free generators, $n \geq 1$. Let $\mathcal{B}_0 \subseteq \mathcal{B}$ be the basis defined in Definition 4.10 and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Let $\mathcal{C} = \{c_1, \dots, c_m\}$, $m \geq 1$, be a subset of \mathcal{B} , and let $P: \mathcal{C} \rightarrow [0, 1]$ be a consistent partial specification of a joint probability distribution on \mathcal{B} . We now consider an arbitrary (yet unknown) joint probability distribution Pr on \mathcal{B} with $Pr|_{\mathcal{C}} = P$. Let the constituent probabilities $Pr(b_i)$, $b_i \in \mathcal{B}_0$, of Pr be denoted by x_i , $i = 1, \dots, 2^n$. Furthermore, let the initially specified probabilities $P(c_i) = Pr(c_i)$, $c_i \in \mathcal{C}$, $i = 1, \dots, m$, be denoted by p_i . From Lemma 4.17 and Lemma 4.19 we obtain the following inhomogeneous system of linear equations:

$$\begin{array}{rcl} d_{1,1}x_1 & + & \dots + d_{1,2^n}x_{2^n} = p_1 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ d_{m,1}x_1 & + & \dots + d_{m,2^n}x_{2^n} = p_m \\ x_1 & + & \dots + x_{2^n} = 1 \end{array}$$

where $d_{ij} = \begin{cases} 0 & \text{if } j \notin \mathcal{J}_{c_i} \\ 1 & \text{if } j \in \mathcal{J}_{c_i} \end{cases}$, $i = 1, \dots, m$, $j = 1, \dots, 2^n$, in which \mathcal{J}_{c_i} is the

index set for $c_i \in \mathcal{C}$. This system of linear equations has the 2^n constituent probabilities of Pr as unknowns. Now, let \mathbf{p} denote the column vector of right-hand sides of this system of linear equations and let \mathbf{x} denote the column vector of unknowns. Furthermore, let \mathbf{D} denote the coefficient matrix of the system. Then, the system of linear equations shown above is equivalent to the following matrix equation:

$$\mathbf{D}\mathbf{x} = \mathbf{p}$$

From now on, we will use this matrix equation to denote the system of linear equations obtained from a partial specification P as described above.

The following lemma states the relation between extensions of a consistent partial specification of a joint probability distribution and solutions to the matrix equation obtained from it.

LEMMA 4.20. *Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$, as defined in Definition 3.7. Let the basis $\mathcal{B}_0 \subseteq \mathcal{B}$ be defined according to Definition 4.10 and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Let $\mathcal{C} \subseteq \mathcal{B}$ and let $P: \mathcal{C} \rightarrow [0, 1]$ be a consistent partial specification of a joint probability distribution on \mathcal{B} . Let $\mathbf{D}\mathbf{x} = \mathbf{p}$ be the matrix equation obtained from P as in the foregoing. Then, the following properties hold:*

- (1) *For any joint probability distribution Pr on \mathcal{B} such that $Pr|_{\mathcal{C}} = P$, we have that the vector \mathbf{x} of constituent probabilities $x_i = Pr(b_i)$, $b_i \in \mathcal{B}_0$, $i = 1, \dots, 2^n$, is a solution to the matrix equation $\mathbf{D}\mathbf{x} = \mathbf{p}$.*
- (2) *For any nonnegative solution vector \mathbf{x} with components x_i , $i = 1, \dots, 2^n$, to the matrix equation $\mathbf{D}\mathbf{x} = \mathbf{p}$, we have that $Pr(b_i) = x_i$, $b_i \in \mathcal{B}_0$, defines a joint probability distribution Pr on \mathcal{B} such that $Pr|_{\mathcal{C}} = P$.*

PROOF. The properties stated in the lemma follow immediately from the Lemmas 4.17 and 4.19. ■

Note that although every joint probability distribution Pr which is an extension of a consistent partial specification P corresponds uniquely with a solution to the matrix equation $\mathbf{D}\mathbf{x} = \mathbf{p}$ obtained from P , not every solution to $\mathbf{D}\mathbf{x} = \mathbf{p}$ corresponds with a 'probabilistic' extension of P : $\mathbf{D}\mathbf{x} = \mathbf{p}$ may have solutions in which at least one of the x_i 's is less than zero.

From Lemma 4.20 we derive a necessary and sufficient condition for a consistent partial specification to be a definition of a joint probability distribution.

COROLLARY 4.21. *Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. Let P be a consistent partial specification of a joint probability distribution on \mathcal{B} and let $D\mathbf{x} = \mathbf{p}$ be the matrix equation obtained from P . Then, P uniquely defines a joint probability distribution on \mathcal{B} if and only if $D\mathbf{x} = \mathbf{p}$ has a unique nonnegative solution.*

Now consider the case in which we are given a consistent partial specification P which can be extended in more than one way to a joint probability distribution on \mathcal{B} , where \mathcal{B} is a Boolean algebra of propositions with n free generators, $n \geq 1$. For making statements concerning probabilities of interest, we can simply select a single 'probabilistic' extension of P and use the selected joint probability distribution for computing the probabilities we are interested in. In the foregoing, we have transformed the problem of finding a joint probability distribution Pr on \mathcal{B} which is an extension of P into the equivalent problem in linear algebra of finding a nonnegative solution to the matrix equation $D\mathbf{x} = \mathbf{p}$ obtained from P . Since P can be extended in more than one way to a joint probability distribution on \mathcal{B} we have that $D\mathbf{x} = \mathbf{p}$ has infinitely many solutions. For the rank r of the coefficient matrix D we have that $r < 2^n$. So, in $D\mathbf{x} = \mathbf{p}$ we have r basic variables and $2^n - r$ free variables. To obtain a particular solution to the matrix equation, we choose the values of the free variables, that is, some of the constituent probabilities, more or less freely although subject to the constraints from the matrix equation and $x_i \geq 0$, $i = 1, \dots, 2^n$; from these values the values of the basic variables can then be computed uniquely.

There are, however, other joint probability distributions on \mathcal{B} respecting the initially given probabilities that are not equal to the one defined by the chosen solution vector: every other nonnegative vector differing from the selected one by a vector in the nullspace of D defines another joint probability distribution on \mathcal{B} which is also an extension of P . It will be evident that the more free variables occur in the matrix equation, the more arbitrary the selected probability distribution will be. The results from using one solution vector for computing probabilities of interest can therefore differ considerably from the results from using another solution vector. Selecting a single, not unique extension of a partial specification of a joint probability distribution to serve as the basis for further computations as sketched in the foregoing, therefore does not render a reliable result.

We abandon the idea of selecting a single extension of a partial specification of a joint probability distribution for further computation: we introduce a method for finding best possible upper and lower bounds on probabilities of interest. The idea of finding bounds on probabilities from a partial specification of a joint probability distribution originated with G. Boole, as well as the idea of obtaining the 'narrowest limits' ([HAIL86], p. 338).

We define the notions of the best upper bound and the best lower bound function relative to a partial specification of a joint probability distribution.

DEFINITION 4.22. Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. Let $\mathcal{C} \subseteq \mathcal{B}$ and let $P: \mathcal{C} \rightarrow [0,1]$ be a consistent partial specification of a joint probability distribution on \mathcal{B} . The function $bub_P: \mathcal{B} \rightarrow [0,1]$ defined by $bub_P(b) = \sup \{Pr(b) \mid Pr \text{ is a joint probability distribution on } \mathcal{B} \text{ such that } Pr|_{\mathcal{C}} = P\}$ for all $b \in \mathcal{B}$, is called the best upper bound function relative to P . The best lower bound function relative to P , denoted by blb_P , is defined symmetrically.

Note that the best upper bound function relative to a partial specification P in general is not a joint probability distribution; of course, the same remark can be made concerning the best lower bound function. Furthermore, for a given $b \in \mathcal{B}$ the length of the interval $[blb_P(b), bub_P(b)]$ expresses the lack of knowledge concerning the probability of the truth of the proposition b .

The two types of bounds are interrelated as stated in the following lemma.

LEMMA 4.23. Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. Let P be a consistent partial specification of a joint probability distribution on \mathcal{B} . Let the functions bub_P and blb_P be defined as above. Then, for each $b \in \mathcal{B}$ we have $bub_P(b) = 1 - blb_P(\neg b)$.

PROOF. From Definition 4.22 we have for each joint probability distribution Pr on \mathcal{B} which is an extension of P and each $b \in \mathcal{B}$, that $Pr(b) \leq bub_P(b)$ and $blb_P(\neg b) \leq Pr(\neg b)$. From $Pr(\neg b) = 1 - Pr(b)$, it follows that $blb_P(\neg b) \leq 1 - Pr(b)$, thus obtaining $Pr(b) \leq 1 - blb_P(\neg b)$, for each $b \in \mathcal{B}$. We therefore have $bub_P(b) \leq 1 - blb_P(\neg b)$. Reversing the argument, we show that $1 - blb_P(\neg b) \leq bub_P(b)$, from which we obtain the property mentioned in the lemma. ■

Let P be a consistent partial specification of a joint probability distribution on a Boolean algebra of propositions \mathcal{B} . The following lemma now states that we can find for each $b \in \mathcal{B}$ a joint probability distribution Pr on \mathcal{B} being an extension of P such that $Pr(b) = bub_P(b)$; again, a similar observation can be made concerning blb_P .

LEMMA 4.24. Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. Let $\mathcal{C} \subseteq \mathcal{B}$ and let $P: \mathcal{C} \rightarrow [0,1]$ be a consistent partial specification of a joint probability distribution on \mathcal{B} . Furthermore, let the functions bub_P and blb_P be defined according to Definition 4.22. Then, for each $b \in \mathcal{B}$ we have $bub_P(b) = \max \{Pr(b) \mid Pr \text{ is a joint probability distribution on } \mathcal{B} \text{ such that } Pr|_{\mathcal{C}} = P\}$. A similar property holds for $blb_P(b)$.

PROOF. The property stated in the lemma will readily be seen using the observation that the Boolean algebra of propositions \mathcal{B} is finite. The lemma has been proven formally by Hailperin, [HAIL65]. ■

On the basis of the properties stated in Lemma 4.24, it can be shown that the problems of finding for a given $b \in \mathcal{B}$ the best upper bound $bub_P(b)$ and the

best lower bound $blb_P(b)$ relative to a consistent partial specification P of a joint probability distribution on \mathcal{B} , are equivalent to the following linear programming problems, respectively:

- (1) maximize $Pr(b)$ subject to $D\mathbf{x} = \mathbf{p}$ and $\mathbf{x} \geq \mathbf{0}$; and
- (2) minimize $Pr(b)$ subject to $D\mathbf{x} = \mathbf{p}$ and $\mathbf{x} \geq \mathbf{0}$,

where $D\mathbf{x} = \mathbf{p}$ is the matrix equation obtained from P . The equivalence will be stated formally in Proposition 4.25. First, we consider case (1) in some detail in order to obtain a more traditional representation of the linear programming problem.

Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$. Let $\mathcal{B}_0 \subseteq \mathcal{B}$ be the basis defined according to Definition 4.10 and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Now let $b \in \mathcal{B}$ and let \mathcal{J}_b be its index set. For each joint probability distribution Pr on \mathcal{B} , we have that

$$Pr(b) = \sum_{i \in \mathcal{J}_b} Pr(b_i) = \sum_{i \in \mathcal{J}_b} x_i$$

Now, let for b constants c_i , $i = 1, \dots, 2^n$, be defined such that

$$c_i = \begin{cases} 0 & \text{if } i \notin \mathcal{J}_b \\ 1 & \text{if } i \in \mathcal{J}_b \end{cases}$$

Then, we have that

$$Pr(b) = \sum_{i=1}^{2^n} c_i x_i$$

So, our aim is to find the best upper bound for this function $\sum_{i=1}^{2^n} c_i x_i$.

We recall that in the matrix equation $D\mathbf{x} = \mathbf{p}$ obtained from a partial specification $P: \mathcal{C} \rightarrow [0, 1]$, $\mathcal{C} \subseteq \mathcal{B}$, D denotes a $(|\mathcal{C}| + 1) \times 2^n$ matrix, \mathbf{x} is the 2^n column vector of constituent probabilities $Pr(b_i)$ and \mathbf{p} is the $|\mathcal{C}| + 1$ column vector of initially given probabilities. The partial problem (1) can therefore be reformulated in the following more traditional representation of a linear programming problem:

$$\text{maximize } \sum_{i=1}^{2^n} c_i x_i$$

subject to

- (i) $\sum_{j=1}^{2^n} d_{ij} x_j = p_i$, for $i = 1, \dots, |\mathcal{C}| + 1$, and
- (ii) $x_j \geq 0$, for $j = 1, \dots, 2^n$,

where the constants $d_{i,j}$ constitute the matrix D . Note that we have added nonnegativity constraints to $D\mathbf{x} = \mathbf{p}$ explicitly to allow for nonnegative solutions only. The linear programming problem (2) can be treated analogously by taking for the objective function $-\sum c_i x_i$.

PROPOSITION 4.25. *Let \mathcal{B} be a Boolean algebra of propositions as defined in Definition 3.7. Let $\mathcal{C} \subseteq \mathcal{B}$ and let $P: \mathcal{C} \rightarrow [0,1]$ be a consistent partial specification of a joint probability distribution on \mathcal{B} . Let $D\mathbf{x} = \mathbf{p}$ be the matrix equation obtained from P . Furthermore, let the functions bub_P and blb_P be defined according to Definition 4.22. Then, for any $b \in \mathcal{B}$ we have that $bub_P(b)$ is equal to the solution of the linear programming problem*

maximize $Pr(b)$

subject to

$$(i) \quad D\mathbf{x} = \mathbf{p}, \text{ and}$$

$$(ii) \quad \mathbf{x} \geq \mathbf{0}.$$

A similar statement can be made concerning $blb_P(b)$.

Note that since computing a best upper bound for a given probability does not 'cut off' solutions from the feasible set of the system of linear constraints, we have that several different objective functions can be maximized independently.

Now consider application of the linear programming approach in a model for handling uncertainty in a knowledge-based system. In short, a domain expert is requested to assess several probabilities. The assessed probabilities are used in the manner described in this section to generate a system of linear constraints. From this system of constraints upper and lower bounds on the probabilities that are of interest to the user of the system are computed. The following example illustrates the idea.

EXAMPLE 4.26. Let $\mathcal{A} = \{a_1, a_2, a_3\}$ and let $\mathcal{B}(a_1, a_2, a_3)$ be the free Boolean algebra generated by \mathcal{A} . Let $\mathcal{C} = \{a_1 \wedge a_2, \neg a_1 \vee a_3, a_2, a_2 \wedge \neg a_3\}$. Note that \mathcal{C} cannot be a basis for a joint probability distribution on $\mathcal{B}(a_1, a_2, a_3)$ since it only contains four elements. Let P be a consistent partial specification defined on \mathcal{C} which can be extended in more than one way to a joint probability distribution on $\mathcal{B}(a_1, a_2, a_3)$. We consider such a 'probabilistic' extension Pr . Suppose that we have the following function values of Pr coinciding with the corresponding initially given function values of P :

$$Pr(a_1 \wedge a_2) = 0.23$$

$$Pr(\neg a_1 \vee a_3) = 0.62$$

$$Pr(a_2) = 0.43$$

$$Pr(a_2 \wedge \neg a_3) = 0.18$$

Now, let the elements of the basis $\mathcal{B}_0 \subseteq \mathcal{B}(a_1, a_2, a_3)$ be enumerated as follows:

$$\begin{aligned}
 b_1 &= a_1 \wedge a_2 \wedge a_3 \\
 b_2 &= \neg a_1 \wedge a_2 \wedge a_3 \\
 b_3 &= a_1 \wedge \neg a_2 \wedge a_3 \\
 b_4 &= a_1 \wedge a_2 \wedge \neg a_3 \\
 b_5 &= \neg a_1 \wedge \neg a_2 \wedge a_3 \\
 b_6 &= \neg a_1 \wedge a_2 \wedge \neg a_3 \\
 b_7 &= a_1 \wedge \neg a_2 \wedge \neg a_3 \\
 b_8 &= \neg a_1 \wedge \neg a_2 \wedge \neg a_3
 \end{aligned}$$

Furthermore, let the constituent probabilities $Pr(b_i)$ be denoted by x_i , $i = 1, \dots, 8$. From P we obtain the following system of linear equations:

$$\begin{aligned}
 x_1 + x_4 &= 0.23 \\
 x_1 + x_2 + x_3 + x_5 + x_6 + x_8 &= 0.62 \\
 x_1 + x_2 + x_4 + x_6 &= 0.43 \\
 x_4 + x_6 &= 0.18 \\
 x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 &= 1
 \end{aligned}$$

We add the constraints

$$x_i \geq 0, i = 1, \dots, 8,$$

explicitly. Now, suppose that we are interested in bounds on the probability of the truth of the atomic proposition a_3 . From Proposition 4.25 we have that the problem of determining the best upper bound for $Pr(a_3)$ is equal to maximizing the objective function

$$x_1 + x_2 + x_3 + x_5$$

subject to the constraints shown above. Applying the simplex method we obtain $bub_P(a_3) = 0.62$. Similarly, we find $blb_P(a_3) = 0.25$. ■

In Section 4.1 we have mentioned that an LP-problem can be solved in polynomial time, that is, polynomial in the size of the problem. Recall that the size of an LP-problem is dependent, among other factors, upon the number of variables it comprises. The specific type of problem discussed in the foregoing has exponentially many variables, that is, exponential in the number of statistical variables discerned in the problem domain. Therefore, these problems cannot be solved in polynomial time; computing bounds on probabilities of interest requires an exponential number of steps.

4.2.2. Dealing with Conditional Probabilities

In the previous subsection we have presented a linear programming method for computing bounds on probabilities of interest from a partial specification of a joint probability distribution. This method has been developed for partial specifications comprising prior probabilities only. In the domains in which expert systems are employed, however, it often is easier to assess or otherwise obtain conditional probabilities than it is to obtain prior ones. Moreover, the user of the system will often be interested in conditional probabilities. We will show that conditional probabilities can be introduced into the linear programming method without requiring much effort.

We first examine the case in which we are initially given some conditional probabilities. Let \mathcal{B} once more be a Boolean algebra of propositions with n free generators, $n \geq 1$, as defined in Definition 3.7. Furthermore, let $\mathcal{B}_0 \subseteq \mathcal{B}$ be the basis as defined in Definition 4.10 and let its elements be enumerated as b_i , $i = 1, \dots, 2^n$. Let P be a consistent partial specification of a joint probability distribution on \mathcal{B} . We consider a joint probability distribution Pr on \mathcal{B} which is an extension of P . Now suppose that an expert has assessed the value $P(c_1|c_2) = Pr(c_1|c_2) = p_0$, where $c_1, c_2 \in \mathcal{B}$ and $0 \leq p_0 \leq 1$, to be taken as a conditional probability. Note that it follows implicitly that $Pr(c_2) \neq 0$. By definition, we have $Pr(c_1|c_2) = \frac{Pr(c_1 \wedge c_2)}{Pr(c_2)}$. From Lemma 4.18 and Lemma 4.19 we have that there exist an index set $\mathcal{J}_{c_1 \wedge c_2}$ for $c_1 \wedge c_2$ such that

$$Pr(c_1 \wedge c_2) = \sum_{i \in \mathcal{J}_{c_1 \wedge c_2}} Pr(b_i)$$

and an index set \mathcal{J}_{c_2} for c_2 such that

$$Pr(c_2) = \sum_{i \in \mathcal{J}_{c_2}} Pr(b_i)$$

where $Pr(b_i)$ are the constituent probabilities of Pr . We therefore have that

$$Pr(c_1|c_2) = \frac{\sum_{i \in \mathcal{J}_{c_1 \wedge c_2}} Pr(b_i)}{\sum_{i \in \mathcal{J}_{c_2}} Pr(b_i)} = p_0$$

It follows that

$$\sum_{i \in \mathcal{J}_{c_1 \wedge c_2}} Pr(b_i) = p_0 \cdot \sum_{i \in \mathcal{J}_{c_2}} Pr(b_i)$$

We now obtain the equation

$$\sum_{i \in \mathcal{J}_{c_1 \wedge c_2}} Pr(b_i) - p_0 \cdot \sum_{i \in \mathcal{J}_{c_2}} Pr(b_i) = 0$$

which is similar in concept to the ones we have encountered before; it can therefore be treated likewise. (Note however that we have to guarantee that $\sum_{i \in \mathcal{I}_2} Pr(b_i) = 0$ is not a solution to the obtained system of linear equations.)

Now consider the case in which we are interested in lower and upper bounds on a conditional probability. From the foregoing discussion, it will be evident that we have a fractional objective function for our problem. Such a problem, called a *fractional linear programming problem*, however, can be reduced to a related 'ordinary' linear programming problem with one additional variable. The following proposition formulated in [HAIL86] but originally due to A. Charnes, states this result.

PROPOSITION 4.27. *The fractional linear programming problem*

$$\begin{aligned} &\text{maximize} \quad \frac{cx}{gx} \\ &\text{subject to} \\ &\quad (i) \quad Dx = p, \text{ and} \\ &\quad (ii) \quad x \geq 0 \end{aligned}$$

is equivalent to the linear programming problem

$$\begin{aligned} &\text{maximize} \quad cy \\ &\text{subject to} \\ &\quad (i) \quad Dy = tp, \\ &\quad (ii) \quad gy = 1, \\ &\quad (iii) \quad y \geq 0, \text{ and} \\ &\quad (iv) \quad t \geq 0. \end{aligned}$$

In Section 4.2.1 we have shown that computing upper and lower bounds on a probability of interest from a consistent partial specification of a joint probability distribution is equivalent to a linear programming problem in which all constraints except the nonnegativity constraints are equalities. The linear programming problems we obtained in the foregoing therefore were in standard form. It will be evident that our method is able to deal with LP-problems in general form as well. Allowing inequalities in our method provides a domain expert with a flexible means for expressing probabilistic information: besides prior and conditional probabilities, he may specify bounds on probabilities instead of point estimates and he may give certain probabilities relative to other ones.

4.2.3. A Note on Inconsistent Partial Specifications

In the foregoing subsections we have dealt with partial specifications that could be extended in at least one way to a joint probability distribution. In Chapter 1, however, we have argued that in an expert system context the 'probabilities' assessed by a domain expert are likely to be inconsistent.

EXAMPLE 4.28. Let $\mathcal{A} = \{a_1, a_2\}$ be a set of atomic propositions and let $\mathcal{B}(a_1, a_2)$ be the Boolean algebra of propositions generated by \mathcal{A} . Let $\mathcal{C} = \{a_1 \wedge a_2, a_1\}$. Now consider the function $P: \mathcal{C} \rightarrow [0, 1]$ defined by

$$P(a_1 \wedge a_2) = 0.34$$

$$P(a_1) = 0.28$$

which is to be taken as a partial specification of a joint probability distribution on $\mathcal{B}(a_1, a_2)$. This function P cannot be extended to a joint probability distribution on $\mathcal{B}(a_1, a_2)$, since in every joint probability distribution Pr on $\mathcal{B}(a_1, a_2)$, for any $x_1, x_2 \in \mathcal{B}$, the property if $x_1 \leq x_2$ then $Pr(x_1) \leq Pr(x_2)$ holds. ■

It will be evident that the more probabilities are initially assessed by a domain expert, the more likely the resulting partial specification is to be inconsistent. In the foregoing example the inconsistency was readily detected. Unfortunately, however, the more probabilities have been assessed, the less evident the inconsistency will be, and therefore the harder to detect and subsequently resolve. In this subsection, we will briefly touch upon the case in which we are given a set of 'probabilities' which is inconsistent in the sense that when the given values are looked upon as values of a partial specification of a joint probability distribution it is not possible to extend this partial specification to an actual joint probability distribution.

We have discussed that the problem of finding an extension of a consistent partial specification P of a joint probability distribution on a Boolean algebra of propositions is equivalent to the problem of finding a nonnegative solution vector x to the system of linear constraints $Dx = p, x \geq 0$, obtained from P . We recall that such a solution vector x is a vector of constituent probabilities. Since in the foregoing we assumed that the initially given probabilities were specified consistently we were guaranteed that the matrix equation $Dx = p$ had at least one such nonnegative solution vector. Now, we have to reckon with the possibility of having an inconsistent partial specification of a joint probability distribution, and we therefore are not guaranteed that the matrix equation obtained has nonnegative solutions; in fact, the matrix equation may have no solution at all or may have only solutions in which at least one of the components is less than zero. It will be evident that inconsistency of the partial specification P corresponds with the system of constraints $Dx = p, x \geq 0$, being infeasible.

The linear programming approach for computing bounds on probabilities of

interest as discussed in Section 4.2.1 cannot be applied when we are given an inconsistent partial specification of a joint probability distribution: for a correct application of our method we need to have a consistent specification. A solution to the problem of having only an inconsistent partial specification at our disposal is to have the expert reassess the function values. This, of course, will be of little practical use since the expert will generally not be able to do any better. We therefore have to obtain a consistent partial specification from the initially given inconsistent one which still reflects the expert's intentions. We emphasize that all methods for doing so can only be viewed as approximation techniques and therefore are apt to be ad hoc. It will be evident that this is an extremely difficult problem and is worth a study of its own. It is noted that in linear algebra, several methods such as *least squares approximations* have been developed for selecting a single vector that best fits an (overdetermined) infeasible system of equations; in statistics a similar method, then called *regression analysis*, is employed. Since these methods aim at selecting a single solution for further computations, we feel that they are not suitable for situations in which only a partial specification of a joint probability distribution is intended. The same more or less holds for methods developed for 'adjusting' subjective probabilities. In further investigating the subject of inconsistency, however, the literature on (the *reconciliation* of) subjective probabilities, such as [LIND79,KAHN82], should not be simply by-passed.

Here, we merely propose an empirical method for obtaining a consistent set of probabilities from an initially given inconsistent partial specification of a joint probability distribution. The general idea of the method is based on the observation that inconsistency of a partial specification may be due to one of the following causes or to a combination of them:

- (1) The normedness is violated, that is, the 'constituent probabilities' computed from the partial specification do not sum up to 1.
- (2) The additivity is violated, that is, the initially specified 'probabilities' are not in correct proportion.

Note that in Example 4.28 (at least) the additivity was violated.

The following lemma applies to an inconsistent partial specification of a joint probability distribution in which only the normedness is violated: the initially specified values behave additively, but the 'constituent probabilities' do not sum up to 1.

LEMMA 4.29. *Let \mathcal{B} be a Boolean algebra of propositions with n free generators, $n \geq 1$, as defined in Definition 3.7. Let $\mathcal{C} \subseteq \mathcal{B}$ and let $P: \mathcal{C} \rightarrow [0,1]$ be an inconsistent partial specification of a joint probability distribution on \mathcal{B} . Let $Dx = p$, $x \geq 0$, be the system of linear constraints obtained from P as described in the preceding subsections; D is a $(|\mathcal{C}| + 1) \times 2^n$ matrix, x is a 2^n column vector and p is a $|\mathcal{C}| + 1$ column vector. We assume that the last row of D is the 1^T row vector. Now, let D^- denote the $|\mathcal{C}| \times 2^n$ matrix obtained from D by omitting its last row; equally, let p^- denote the $|\mathcal{C}|$ column vector obtained*

from \mathbf{p} by deleting its last component. If the system of linear constraints $\mathbf{D}^-\mathbf{x} = \mathbf{p}^-$, $\mathbf{x} \geq \mathbf{0}$, has a solution and $\mathbf{p}^- \neq \mathbf{0}$, then there exists a scalar $k > 0$ such that the system of constraints $\mathbf{D}\mathbf{x} = \begin{bmatrix} k\mathbf{p}^- \\ 1 \end{bmatrix}$, $\mathbf{x} \geq \mathbf{0}$, has a solution.

PROOF. Suppose that the system of linear constraints $\mathbf{D}^-\mathbf{x} = \mathbf{p}^-$, $\mathbf{x} \geq \mathbf{0}$, has at least one solution. We consider such a solution vector \mathbf{x}' with components x'_j , $j = 1, \dots, 2^n$. For this vector \mathbf{x}' we evidently have

$$\sum_{j=1}^{2^n} d_{ij}x'_j = p_i$$

$i = 1, \dots, |\mathcal{C}|$, where d_{ij} constitute \mathbf{D}^- and p_i constitute the vector \mathbf{p}^- . From the system of linear constraints $\mathbf{D}\mathbf{x} = \mathbf{p}$, $\mathbf{x} \geq \mathbf{0}$, not having a solution, we have that either $\sum x'_j < 1$ or $\sum x'_j > 1$. Furthermore, from $\mathbf{p}^- \neq \mathbf{0}$ it follows that $\sum x'_j > 0$. Now, let $y'_j = \frac{x'_j}{\sum_{j=1}^{2^n} x'_j}$. Then, we have

$$\sum_{j=1}^{2^n} d_{ij}y'_j = \frac{p_i}{\sum_{j=1}^{2^n} x'_j}$$

$i = 1, \dots, |\mathcal{C}|$. From \mathbf{x}' being a solution to the system of constraints $\mathbf{D}^-\mathbf{x} = \mathbf{p}^-$, $\mathbf{x} \geq \mathbf{0}$, we have that \mathbf{y}' with components y'_j is a solution to the system of linear constraints $\mathbf{D}^-\mathbf{x} = k\mathbf{p}^-$, $\mathbf{x} \geq \mathbf{0}$, where $k = \frac{1}{\sum_{j=1}^{2^n} x'_j}$. It will

be evident that $\sum_{j=1}^{2^n} y'_j = 1$. Therefore, the vector \mathbf{y}' is a solution to the system of linear constraints $\mathbf{D}\mathbf{x} = \begin{bmatrix} k\mathbf{p}^- \\ 1 \end{bmatrix}$, $\mathbf{x} \geq \mathbf{0}$. ■

The basic idea of Lemma 4.29 is that, interpreted in 2^n -dimensional space, the convex polyhedron F being the feasible set of the system of linear constraints $\mathbf{D}^-\mathbf{x} = \mathbf{p}^-$, $\mathbf{x} \geq \mathbf{0}$, is moved along the \mathbf{p}^- vector towards the origin or just away from the origin dependent upon whether $\sum x_i > 1$ or $\sum x_i < 1$, so that the intersection of the shifted polyhedron and the hyperplane $\sum x_i = 1$ is not empty. It will be evident that there exist many scalars having the property mentioned in the proposition, obtained from different points in the original convex polyhedron F .

The method we propose for obtaining a consistent set of probabilities from an inconsistent partial specification of a joint probability distribution which does not behave additively, is based on the idea of allowing an expert to make a certain mistake in his assessments. For each probability the expert has assessed, we add two inequalities to the system of linear constraints instead of

one equality. Let $0 < m < 1$ be a small constant value representing the margin we allow the expert to be mistaken in his assessments. When an expert has specified the value $Pr(c) = \sum d_{ij}x_j = p_0$, $0 \leq p_0 \leq 1$, we add the following inequalities to the system of constraints:

$$(1) \quad \sum_{j=1}^{2^n} d_{ij}x_j \leq p_0^+ \text{ where } p_0^+ = \begin{cases} p_0 + m & \text{if } p_0 + m < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$(2) \quad \sum_{j=1}^{2^n} d_{ij}x_j \geq p_0^- \text{ where } p_0^- = \begin{cases} p_0 - m & \text{if } p_0 - m > 0 \\ 0 & \text{otherwise} \end{cases}$$

Notice that instead of a hyperplane we have specified a 'band' in 2^n -dimensional space. We consider the system of linear constraints that is obtained this way from the expert's assessments. If this system is still infeasible, then the assessments cannot be used and the expert has to reassess the probabilities. If the resulting system of constraints however has at least one feasible solution, then the equation $\sum x_i = 1$ is added to the system (after applying Lemma 4.29, if necessary). We then proceed with the thus obtained system of linear constraints. Note that in this approach the equalities from the original matrix equation are treated as being equally trustworthy. If the domain expert is more certain of some of his assessments than of the other ones, however, we can attach for each constraint a weighting factor to the margin m we allow the expert to be mistaken thus obtaining a constraint-specific margin determining the width of the specified band.

4.3. PARTIAL QUANTIFICATION OF A BELIEF NETWORK

In the preceding section we have presented a linear programming method for computing bounds on probabilities of interest from a consistent partial specification of a joint probability distribution. The initially assessed probabilities were viewed as defining constraints on an unknown probability distribution. We assumed that no independency relationships existed between the statistical variables discerned in the problem domain. For the linear programming method to be applicable to a partially quantified belief network, however, it has to be extended with an additional method for representing and exploiting independency relationships between the variables. Note that representing independency relationships in a straightforward manner yields nonlinear equations and therefore is not suitable for our purposes.

We recall from the introduction to this chapter that a partially quantified belief network was meant to consist of an acyclic directed graph representing the statistical variables discerned in the problem domain and their independency relationships, and an associated partial specification of a joint probability distribution. In this section we will present a method for computing bounds on probabilities of interest from such a partially quantified belief network in which the independency relationships portrayed by its graph

are exploited. Our method builds on the work on belief networks by S.L. Lauritzen and D.J. Spiegelhalter as discussed in Chapter 3. For ease of exposition we will assume that the graphical part of a given partially quantified belief network has been transformed into a decomposable graph as demonstrated in Section 3.4; we will return to this simplifying assumption at the end of this section. In the mean time, we will show that we can take advantage of the topology of a decomposable graph by observing that between the statistical variables all occurring in the same clique of the graph no independency relationships exist and that the cliques are interrelated only through their intersections. In order to be able to exploit these properties, we further assume that all initially given probabilities are prior ones and local to the cliques of the graph G : we are given probabilities $P(c_V)$ where each c_V is a configuration of a set of vertices V such that there is at least one clique Cl_i in G with $V \subseteq V(Cl_i)$. Recall from the previous section that the restriction to prior probabilities is not an essential one; the restriction to local probabilities, however, is essential.

The initially given probabilities being local to the cliques of the decomposable graph G of a partially quantified belief network now allows us to apply many of the notions introduced in the preceding section separately to the cliques of G and their associated marginal distributions. We begin by using the definition of a partial specification of a joint probability distribution to apply to marginal distributions.

DEFINITION 4.30. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$. For each clique $Cl_i \in Cl(G)$, let $\mathcal{B}(Cl_i) \subseteq \mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_j \mid V_j \in V(Cl_i)\}$. A partial specification of a marginal distribution on $\mathcal{B}(Cl_i)$ is a total function $m_{Cl_i}: \mathcal{C}_i \rightarrow [0, 1]$ where $\mathcal{C}_i \subseteq \mathcal{B}(Cl_i)$.*

We now are able to define the notion of a partially quantified belief network more formally; note that for the moment we assume that the graphical part of a partially quantified belief network is a decomposable graph.

DEFINITION 4.31. *A partially quantified belief network is a tuple $B = (G, M)$ such that*

- (1) *G is a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, and*
- (2) *$M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ is a set of partial specifications of marginal distributions m_{Cl_i} on $\mathcal{B}(Cl_i)$, where $\mathcal{B}(Cl_i)$ is the free Boolean algebra associated with clique Cl_i as indicated in Definition 4.30.*

We recall from Definition 4.8 that we defined a partial specification of a joint probability distribution as being consistent if it could be extended in at least

one way to an actual joint probability distribution. In Definition 4.30 we have used the definition of a partial specification of a joint probability distribution to apply to marginal distributions; we now take the notion of consistency to apply to a partial specification of a marginal distribution.

DEFINITION 4.32. Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$, and for each clique $Cl_i \in Cl(G)$, let $\mathcal{B}(Cl_i) \subseteq \mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra associated with Cl_i as in Definition 4.30. A partial specification of a marginal distribution $m_{Cl_i}: \mathcal{C}_i \rightarrow [0, 1]$, $\mathcal{C}_i \subseteq \mathcal{B}(Cl_i)$, is consistent if there exists at least one marginal distribution μ_{Cl_i} on $\mathcal{B}(Cl_i)$ such that $\mu_{Cl_i}|_{\mathcal{C}_i} = m_{Cl_i}$; otherwise, m_{Cl_i} is said to be inconsistent.

Now observe that between the statistical variables occurring in one and the same clique of a decomposable graph no independency relationships exist. This observation and the analogy between the notions of a consistent partial specification of a joint probability distribution and a consistent partial specification of a marginal distribution suggest that we may apply the linear programming method presented in the preceding section separately to each of the partial specifications associated with the cliques of the graph. It will be evident, however, that even if all partial specifications of marginal distributions associated with the cliques of the decomposable graph have been specified consistently and therefore can be extended separately to marginal distributions, they might not give rise to a joint probability distribution respecting the independency relationships shown in the graph. Therefore, we define two notions of consistency for a set of partial specifications of marginal distributions.

DEFINITION 4.33. Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$, and for each clique $Cl_i \in Cl(G)$, let $\mathcal{B}(Cl_i) \subseteq \mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_j \mid V_j \in V(Cl_i)\}$. Let $M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ be a set of partial specifications of marginal distributions such that $B = (G, M)$ is a partially quantified belief network as defined in Definition 4.31.

- (1) M is called *locally consistent* if each partial specification $m_{Cl_i} \in M$, $i = 1, \dots, m$, is consistent.
- (2) M is called *globally consistent* if there exists a set $\bar{M} = \{\mu_{Cl_i} \mid \mu_{Cl_i}: \mathcal{B}(Cl_i) \rightarrow [0, 1]\}$ of marginal distributions μ_{Cl_i} on $\mathcal{B}(Cl_i)$ such that for each clique $Cl_i \in Cl(G)$, μ_{Cl_i} is an extension of $m_{Cl_i} \in M$, and furthermore that $\mu_{Cl_i}(C_{V(Cl_i) \cap V(Cl_j)}) = \mu_{Cl_j}(C_{V(Cl_i) \cap V(Cl_j)})$ for each pair of cliques $Cl_i, Cl_j \in Cl(G)$ with $V(Cl_i) \cap V(Cl_j) \neq \emptyset$; such a set \bar{M} is called a *global extension* of M .

For a partial specification of a joint probability distribution we define an additional notion of inconsistency related to a decomposable graph.

DEFINITION 4.34. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$. Let P be a partial specification of a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$. We say that P is consistent with respect to G if P can be extended in at least one way to a joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_n)$ such that Pr is decomposable relative to G .*

Let $B = (G, M)$ be a partially quantified belief network. Now, consider the partial specification P of a joint probability distribution defined by the set M of partial specifications of marginal distributions, simply by taking the function values of the partial specifications from M as function values of P . The following lemma states that global consistency of the set M is a necessary and sufficient condition for P being consistent with respect to the decomposable graph G .

LEMMA 4.35. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$, and for each clique $Cl_i \in Cl(G)$, let $\mathcal{B}(Cl_i)$ be the free Boolean algebra generated by $\{v_j \mid V_j \in V(Cl_i)\}$. Let $M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ be a set of partial specifications of marginal distributions such that $B = (G, M)$ is a partially quantified belief network. Now, let $\mathcal{C} \subseteq \mathcal{B}(v_1, \dots, v_n)$ be such that $\mathcal{C} = \bigcup \{\mathcal{C}_i \mid m_{Cl_i}: \mathcal{C}_i \rightarrow [0, 1], m_{Cl_i} \in M, \mathcal{C}_i \subseteq \mathcal{B}(Cl_i)\}$. Furthermore, let $P: \mathcal{C} \rightarrow [0, 1]$ be the partial specification of a joint probability distribution on $\mathcal{B}(v_1, \dots, v_n)$ defined by $P(c) = m_{Cl_i}(c)$ for each $c \in \mathcal{C}_i$, $i = 1, \dots, m$. Then, M is globally consistent if and only if P is consistent with respect to G .*

PROOF.

\Rightarrow Let $M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ be a globally consistent set of partial specifications of marginal distributions $m_{Cl_i}: \mathcal{C}_i \rightarrow [0, 1]$, $\mathcal{C}_i \subseteq \mathcal{B}(Cl_i)$, $i = 1, \dots, m$. We have from Definition 4.33 that there exists a set $M = \{\mu_{Cl_i} \mid \mu_{Cl_i}: \mathcal{B}(Cl_i) \rightarrow [0, 1]\}$ of marginal distributions μ_{Cl_i} on $\mathcal{B}(Cl_i)$ such that for each clique $Cl_i \in Cl(G)$, μ_{Cl_i} is an extension of m_{Cl_i} , and furthermore that $\mu_{Cl_i}(C_{V(Cl_i) \cap V(Cl_j)}) = \mu_{Cl_j}(C_{V(Cl_i) \cap V(Cl_j)})$ for each pair of cliques $Cl_i, Cl_j \in Cl(G)$ with $V(Cl_i) \cap V(Cl_j) \neq \emptyset$. From μ_{Cl_i} being an extension of m_{Cl_i} , we have for each $c \in \mathcal{C}_i$ that $\mu_{Cl_i}(c) = m_{Cl_i}(c)$, $i = 1, \dots, m$. From Lemma 3.40 we have that the set M defines a joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_n)$ such that Pr is decomposable relative to G . We have $Pr(c) = \mu_{Cl_i}(c)$ for each $c \in \mathcal{C}_i$, $i = 1, \dots, m$. So, $Pr(c) = m_{Cl_i}(c) = P(c)$ for each $c \in \mathcal{C}_i$, $i = 1, \dots, m$. It follows that $Pr|_{\mathcal{C}} = P$. We have by definition that P is consistent with respect to G .

← Let P be a partial specification of a joint probability distribution which is consistent with respect to G . By definition we have that there exists a joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_n)$ being an extension of P which is decomposable relative to G . It follows that $Pr(c) = P(c)$ for each $c \in \mathcal{C}$. From Proposition 3.38 we furthermore have that Pr can be written in terms of marginal distributions μ_{Cl_i} on $\mathcal{B}(Cl_i)$, $i = 1, \dots, m$, that is, Pr is defined by a set $M = \{\mu_{Cl_i} \mid \mu_{Cl_i}: \mathcal{B}(Cl_i) \rightarrow [0, 1]\}$ of marginal distributions such that for each $Cl_i, Cl_j \in Cl(G)$ with $V(Cl_i) \cap V(Cl_j) \neq \emptyset$ we have $\mu_{Cl_i}(C_{V(Cl_i) \cap V(Cl_j)}) = \mu_{Cl_j}(C_{V(Cl_i) \cap V(Cl_j)})$. Recall that we have assumed that the initially given probabilities are local to the cliques of G . So, for each $c \in \mathcal{C}$, we have that there exists an index i such that $c \in \mathcal{C}_i$. For all $c \in \mathcal{C}_i$, $i = 1, \dots, m$, we therefore have $Pr(c) = \mu_{Cl_i}(c)$. So, $\mu_{Cl_i}(c) = P(c) = m_{Cl_i}(c)$ for each $c \in \mathcal{C}_i$. It follows that $\mu_{Cl_i}|_{\mathcal{C}_i} = m_{Cl_i}$, for each $Cl_i \in Cl(G)$, $i = 1, \dots, m$. So, M is a global extension of M . From Definition 4.33 we have that M is globally consistent. ■

The following corollary follows straight from the proof of the preceding lemma.

COROLLARY 4.36. *Let $B = (G, M)$ be a partially quantified belief network, where G is a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$. Furthermore, let P be the partial specification of a joint probability distribution defined by M as in Lemma 4.35. Then, each global extension M of M defines a joint probability distribution Pr on $\mathcal{B}(v_1, \dots, v_n)$ such that Pr is an extension of P and Pr is decomposable relative to G , and vice versa.*

Let $B = (G, M)$ once more be a partially quantified belief network. We have suggested before that we can apply the linear programming method separately to each of the partial specifications of marginal distributions associated with the cliques of the decomposable graph G . To this end, we now define for each clique Cl_i of G a set of constituent probabilities of a yet unknown marginal distribution μ_{Cl_i} in the manner described in the previous section. From the partial specification m_{Cl_i} associated with clique Cl_i we then obtain an appropriate system of linear constraints with the constituent probabilities as unknowns just as we have done before. From our observations from Section 4.2 we have that each nonnegative solution to such a system of constraints defines a marginal distribution which is an extension of the corresponding partial specification m_{Cl_i} . The separate systems of constraints are subsequently combined into one large system of linear constraints. Note that the separate

systems of constraints corresponding with the different cliques do not have any variables, that is, constituent probabilities, in common and therefore are not interrelated. We now have to guarantee that every nonnegative solution to the thus obtained system of constraints defines an extension of the initially given probabilities to an actual joint probability distribution respecting the independency relationships portrayed by G . We have from Lemma 4.35 that in order to guarantee this, it suffices to augment the system of constraints with some additional equations expressing that the set M of partial specifications of marginal distributions has to be globally consistent. Henceforth, such additional equations will be called *consistency equations*. Note that these consistency equations will specify constituent probabilities from more than one clique. The following example illustrates the basic idea.

EXAMPLE 4.37. Let $G = (V(G), E(G))$ be the decomposable graph shown in Figure 4.1.

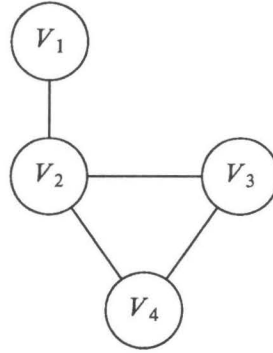


FIGURE 4.1. A decomposable graph G .

Let $\mathcal{B}(v_1, \dots, v_4)$ be the Boolean algebra of propositions associated with G . The graph G has only two cliques: the clique Cl_1 with the vertex set $V(Cl_1) = \{V_1, V_2\}$ and the clique Cl_2 with the vertex set $V(Cl_2) = \{V_2, V_3, V_4\}$. Let $\mathcal{B}(v_1, v_2)$ and $\mathcal{B}(v_2, v_3, v_4)$ be the free Boolean algebras associated with the cliques Cl_1 and Cl_2 , respectively. Now, let $M = \{m_{Cl_1}, m_{Cl_2}\}$ be a set of partial specifications of marginal distributions such that $B = (G, M)$ is a partially quantified belief network. For clique Cl_1 we define the following constituent probabilities:

$$\begin{aligned} x_1^1 &= \mu_{Cl_1}(v_1 \wedge v_2) & x_3^1 &= \mu_{Cl_1}(v_1 \wedge \neg v_2) \\ x_2^1 &= \mu_{Cl_1}(\neg v_1 \wedge v_2) & x_4^1 &= \mu_{Cl_1}(\neg v_1 \wedge \neg v_2) \end{aligned}$$

where μ_{Cl_1} is a yet unknown marginal distribution on $\mathcal{B}(v_1, v_2)$. For clique Cl_2 we define the following eight constituent probabilities:

$$\begin{aligned} x_1^2 &= \mu_{Cl_2}(v_2 \wedge v_3 \wedge v_4) & x_5^2 &= \mu_{Cl_2}(\neg v_2 \wedge \neg v_3 \wedge v_4) \\ x_2^2 &= \mu_{Cl_2}(\neg v_2 \wedge v_3 \wedge v_4) & x_6^2 &= \mu_{Cl_2}(\neg v_2 \wedge v_3 \wedge \neg v_4) \\ x_3^2 &= \mu_{Cl_2}(v_2 \wedge \neg v_3 \wedge v_4) & x_7^2 &= \mu_{Cl_2}(v_2 \wedge \neg v_3 \wedge \neg v_4) \\ x_4^2 &= \mu_{Cl_2}(v_2 \wedge v_3 \wedge \neg v_4) & x_8^2 &= \mu_{Cl_2}(\neg v_2 \wedge \neg v_3 \wedge \neg v_4) \end{aligned}$$

where μ_{Cl_2} is an unknown marginal distribution on $\mathcal{B}(v_2, v_3, v_4)$. From the partial specifications $m_{Cl_1} \in M$ and $m_{Cl_2} \in M$ corresponding with the cliques Cl_1 and Cl_2 , respectively, we obtain two systems of linear constraints in the manner described in the previous subsection. These systems comprise the variables x_i^1 , $i = 1, \dots, 4$, and x_i^2 , $i = 1, \dots, 8$, respectively. It will be evident that these systems do not have any variables in common. The two systems of constraints are combined into one large system of constraints. We now have to augment the thus obtained system of constraints with consistency equations expressing that the set M has to be globally consistent. We therefore add equations expressing that any global extension $M = \{\mu_{Cl_1}, \mu_{Cl_2}\}$ of M has to satisfy the property $\mu_{Cl_1}(V_2) = \mu_{Cl_2}(V_2)$. We obtain the following equations:

$$\begin{aligned} x_1^1 + x_2^1 &= x_1^2 + x_3^2 + x_4^2 + x_7^2 \\ x_3^1 + x_4^1 &= x_2^2 + x_5^2 + x_6^2 + x_8^2 \end{aligned}$$

(Note that actually it suffices to specify only one of these equations). ■

In the way demonstrated in the preceding example, we obtain consistency equations for each nonempty clique intersection. We then have obtained a system of linear constraints having the form shown in Figure 4.2; a system of constraints of this form is called an *angular* system of constraints.

Definition 4.38 introduces the notion of a clique-incidence graph showing all nonempty clique intersections for a given decomposable graph; it is an undirected graph in which the cliques of the original graph are taken as vertices and in which occurrences of nonempty clique intersections are represented by edges.

DEFINITION 4.38. Let $G = (V(G), E(G))$ be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, be the clique set of G . The clique-incidence graph I_G associated with G is the undirected graph $I_G = (V(I_G), E(I_G))$ where $V(I_G) = \{Cl_i \mid Cl_i \in Cl(G)\}$ and $E(I_G) = \{(Cl_i, Cl_j) \mid V(Cl_i) \cap V(Cl_j) \neq \emptyset, Cl_i, Cl_j \in Cl(G), i < j\}$.

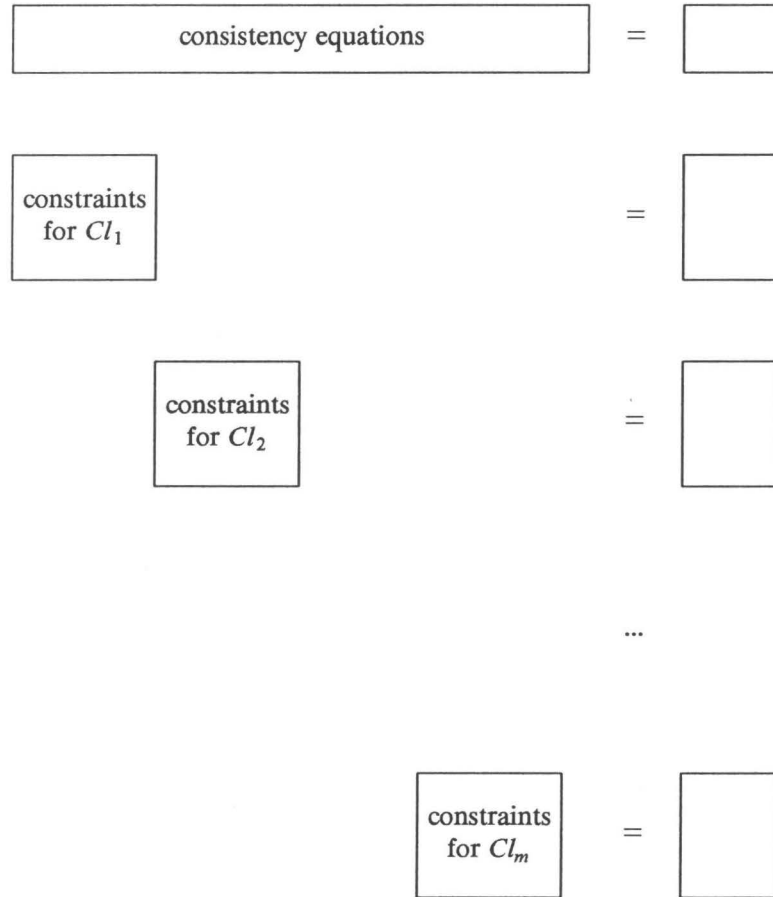


FIGURE 4.2. Linear constraints from a partially quantified belief network.

EXAMPLE 4.39. Consider the decomposable graph G from Figure 3.5(b) once more. The clique-incidence graph I_G associated with G is shown in Figure 4.3; for simplicity's sake, we have identified a clique with its vertex set. ■

Recall that from a set M of partial specifications of marginal distributions associated with a decomposable graph G , we have obtained an angular system of linear constraints of the form shown in Figure 4.2. From the foregoing discussion, it will be evident that this system of constraints comprises consistency equations for all edges of the clique-incidence graph associated with G . The following example, however, shows that the system of constraints thus obtained specifies many redundant equations.

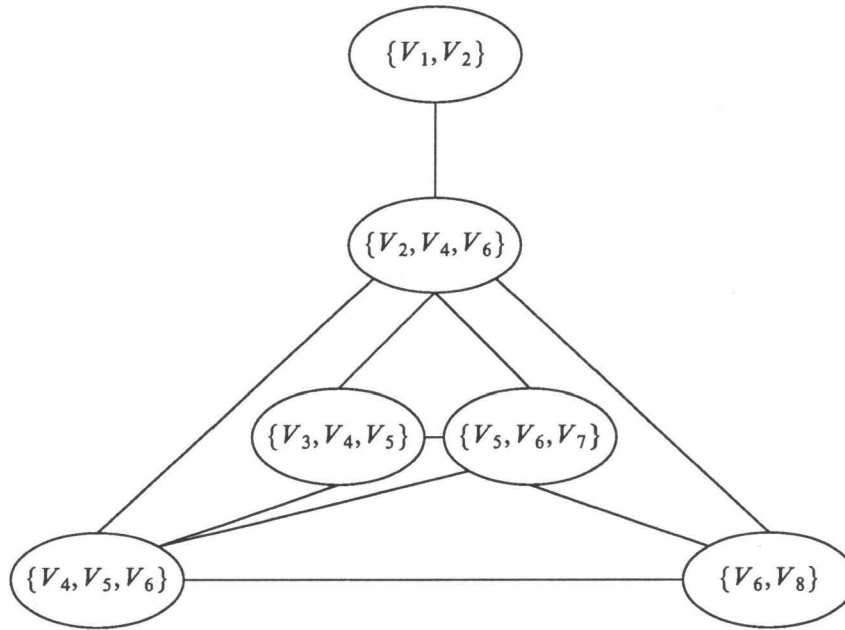


FIGURE 4.3. The clique-incidence graph I_G associated with G .

EXAMPLE 4.40. Consider the decomposable graph G from Figure 3.5(b) and its clique-incidence graph I_G as shown in Figure 4.3 once more. We now consider the cliques with the following vertex sets in isolation:

$$V(Cl_1) = \{V_2, V_4, V_6\}$$

$$V(Cl_2) = \{V_5, V_6, V_7\}$$

$$V(Cl_3) = \{V_6, V_8\}$$

Note that these cliques share the variable V_6 . From the edge (Cl_1, Cl_2) we obtain two consistency equations expressing that any two marginal distributions μ_{Cl_1} and μ_{Cl_2} associated with the cliques Cl_1 and Cl_2 , respectively, have to satisfy

$$(1) \quad \mu_{Cl_1}(v_6) = \mu_{Cl_2}(v_6)$$

$$(2) \quad \mu_{Cl_1}(\neg v_6) = \mu_{Cl_2}(\neg v_6)$$

From the edges (Cl_2, Cl_3) and (Cl_1, Cl_3) we obtain consistency equations expressing that:

$$(3) \quad \mu_{Cl_2}(v_6) = \mu_{Cl_3}(v_6)$$

$$(4) \quad \mu_{Cl_2}(\neg v_6) = \mu_{Cl_3}(\neg v_6)$$

$$(5) \quad \mu_{Cl_1}(v_6) = \mu_{Cl_3}(v_6)$$

$$(6) \quad \mu_{Cl_1}(\neg v_6) = \mu_{Cl_3}(\neg v_6)$$

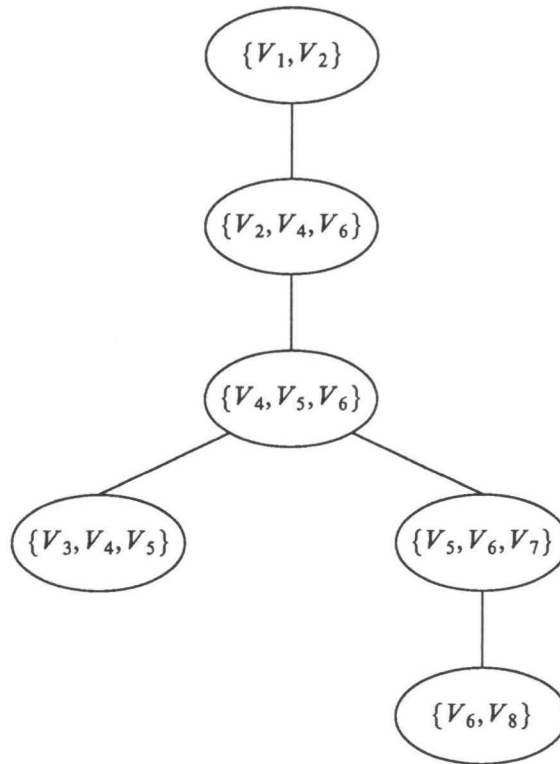
From the transitivity of the equality, however, we for example have that equation (5) immediately follows from the equations (1) and (3), and that equation (6) follows from (2) and (4). In fact, it will suffice to specify consistency equations for only two of the three edges (Cl_1, Cl_2) , (Cl_2, Cl_3) and (Cl_1, Cl_3) . ■

In Definition 4.41 we introduce the notion of a clique tree of a decomposable graph. After stating some properties of such a clique tree we will show in Lemma 4.45 that for guaranteeing global consistency of a set of partial specifications of marginal distributions it suffices to obtain consistency equations from such a clique tree only instead of from the entire clique-incidence graph.

DEFINITION 4.41. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, be the clique set of G . A clique tree of G is a tree $T_G = (V(T_G), E(T_G))$ where $V(T_G) = \{Cl_i \mid Cl_i \in Cl(G)\}$ and where $E(T_G)$ has the following property: for each pair of distinct cliques $Cl_i, Cl_j \in Cl(G)$ with $V(Cl_i) \cap V(Cl_j) \neq \emptyset$, the (unique) path in T_G from Cl_i to Cl_j is a sequence of vertices Cl_k such that $V(Cl_i) \cap V(Cl_j) \subseteq V(Cl_k)$.*

The notion of a clique tree is frequently used in the recent literature on belief networks. Our notion of a clique tree for example is equal to the notion of a *junction tree* as defined by F.V. Jensen in [JENS88a, JENS88b] and similar to the notion of a *join tree* as discussed by J. Pearl, [PEAR88]. In [DEMP88], A.P. Dempster and A. Kong introduce a so-called *tree of cliques* which is analogous to our clique tree as well. Not only in the research area of plausible reasoning are clique trees and their related properties encountered. The notion of a clique tree is of major importance to the theory of acyclic databases [MAIE83]; there it is again called a join tree. In their work on constraint satisfaction problems, R. Dechter and J. Pearl employ related techniques as well, see for example [DECH87].

EXAMPLE 4.42. Figure 4.4 shows a clique tree for our running example. Note that this clique tree is not the only one for our graph. In fact, a decomposable graph may have more than one clique tree. ■

FIGURE 4.4. A clique tree T_G of G .

The following lemma will be evident.

LEMMA 4.43. *Let G be a decomposable graph and let I_G be its associated clique-incidence graph as defined in Definition 4.38. Then, each clique tree T_G of G is a spanning tree of I_G .*

The reverse property does not hold, that is, not every spanning tree of the clique-incidence graph is a clique tree of the original decomposable graph.

The following lemma provides a method for constructing a clique tree of a given decomposable graph. For further details, see [PEAR88].

LEMMA 4.44. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$. Let $Cl(G)$ be the set of cliques in G , numbered Cl_1, \dots, Cl_m , $m \geq 1$, according to an ordering \hat{i} having the running intersection property. The following algorithm yields a clique tree of G :*

1. Start with the tree consisting of Cl_1 only.
2. Subsequently, add the remaining cliques to the tree in increasing order according to \hat{i} such that Cl_i is connected to a clique Cl_j , $j < i$, sharing the highest number of vertices with Cl_i .

The notion of a clique tree is already implicitly present in the work of S.L. Lauritzen and D.J. Spiegelhalter. Recall from Chapter 3 that in their scheme for evidence propagation in a belief network, for each new piece of evidence the vertices and the cliques of the decomposable graph G of the (transformed) network are ordered anew using Algorithm 3.31 and Definition 3.33, starting with a clique containing the observed vertex; the obtained ordering \hat{i} of the cliques then is taken as the order in which the evidence is propagated through the network. From Lemma 4.44 we have that the ordering \hat{i} may be used for constructing a clique tree of the graph G ; propagating a piece of evidence through the belief network as prescribed by Lauritzen and Spiegelhalter now amounts to propagating the evidence through the clique tree obtained from \hat{i} , starting with its root. Now, note that the clique tree can be exploited to render recomputing \hat{i} for new evidence unnecessary: for propagating a new piece of evidence we simply take an appropriate vertex of the clique tree as its new root.

Before proceeding we introduce some new notational convention. Let $B = (G, M)$ be a partially quantified belief network. Recall that for each clique Cl_i , $i = 1, \dots, m$, $m \geq 1$, of the decomposable graph G , we have defined new constituent probabilities; the vector of constituent probabilities for clique Cl_i will from now on be denoted as x_i . Analogous to the notational convention introduced in the previous section, we use $D_i x_i = m_i$ to denote the system of linear constraints obtained from the partial specification of a marginal distribution $m_{Cl_i} \in M$ associated with clique Cl_i . The nonnegativity constraints for Cl_i are expressed as $x_i \geq 0$, $i = 1, \dots, m$. The separate systems of constraints corresponding with the cliques of G are combined into one large system of linear constraints; this system will be denoted as $Dx = m$, $x \geq 0$, where

$$D = \begin{bmatrix} D_1 & 0 & \dots & 0 \\ 0 & D_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D_m \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \quad \text{and} \quad m = \begin{bmatrix} m_1 \\ \vdots \\ m_m \end{bmatrix}$$

Furthermore, recall that this system of constraints is extended with consistency equations expressing that the set M of partial specifications of marginal distributions has to be globally consistent. It will be evident that these

consistency equations each involve variables from two cliques only. In the sequel, the system of consistency equations for the nonempty intersection of two distinct cliques Cl_i and Cl_j will be denoted by $T_{i,j}x_i - T_{j,i}x_j = \mathbf{0}$. The system of consistency equations obtained from an entire clique tree of the decomposable graph G will be denoted by $Tx = \mathbf{0}$, where T is a block matrix with two nonzero blocks per row. We will call the system of constraints $Dx = m$, $Tx = \mathbf{0}$, $x \geq \mathbf{0}$, the *joint* system of constraints. From now on, we assume that this joint system of constraints is feasible. To conclude, a solution vector $x = (x_{1,1}, \dots, x_{1,k_1}, \dots, x_{m,1}, \dots, x_{m,k_m})$, $k_i \geq 1$, $i = 1, \dots, m$, will often be written as $x = (x_1, \dots, x_m)$ where $x_i = (x_{i,1}, \dots, x_{i,k_i})$ is the part of the solution vector x corresponding with the subsystem of constraints $D_i x_i = m_i$, $x_i \geq \mathbf{0}$.

The following lemma now states that for obtaining consistency equations a clique tree of the decomposable graph G suffices.

LEMMA 4.45. *Let $B = (G, M)$ be a partially quantified belief network defined as in Definition 4.31. Let $Dx = m$ be the system of linear constraints obtained from M as described in the foregoing. Now, let I_G be the clique-incidence graph associated with G , and let $Jx = \mathbf{0}$ be the system of consistency equations obtained from I_G ; we use F_I to denote the feasible set of the system of constraints $Dx = m$, $Jx = \mathbf{0}$, $x \geq \mathbf{0}$. Furthermore, let T_G be a clique tree of G , and let $Tx = \mathbf{0}$ be the system of consistency equations obtained from T_G ; we use F_T to denote the feasible set of the system of constraints $Dx = m$, $Tx = \mathbf{0}$, $x \geq \mathbf{0}$. Then, $F_I = F_T$.*

PROOF. From Lemma 4.43 we have that T_G is a spanning tree of I_G . It follows that $Tx = \mathbf{0}$ is a subsystem of $Jx = \mathbf{0}$. So, we have $F_I \subseteq F_T$. It now suffices to show that the addition of an arbitrary edge from I_G to T_G does not yield further restrictions on the set F_T : we have to show that the consistency equations obtained from the new edge follow from the system of constraints $Tx = \mathbf{0}$.

Consider part of a clique tree T_G , as shown in Figure 4.5(a).

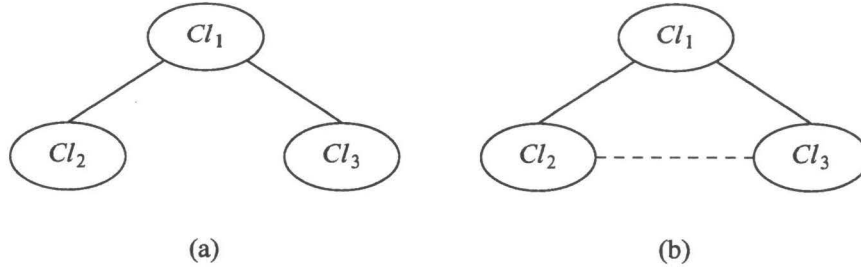


FIGURE 4.5. The addition of an edge from I_G to T_G .

From the edges (Cl_1, Cl_2) and (Cl_1, Cl_3) we obtain the following consistency equations (for ease of exposition we express the consistency equations in terms of marginal distributions instead of in terms of the constituent probabilities of these marginal distributions):

- (1) $\mu_{Cl_1}(c_{V(Cl_1) \cap V(Cl_2)}) = \mu_{Cl_2}(c_{V(Cl_1) \cap V(Cl_2)}),$ for each configuration $c_{V(Cl_1) \cap V(Cl_2)}$ of $V(Cl_1) \cap V(Cl_2)$, and
- (2) $\mu_{Cl_1}(c_{V(Cl_1) \cap V(Cl_3)}) = \mu_{Cl_3}(c_{V(Cl_1) \cap V(Cl_3)}),$ for each configuration $c_{V(Cl_1) \cap V(Cl_3)}$ of $V(Cl_1) \cap V(Cl_3)$.

Now suppose that I_G contains the edge (Cl_2, Cl_3) . From T_G being a spanning tree of I_G we have that the addition of an arbitrary edge from I_G which is not already in T_G to T_G yields a cycle. Addition of (Cl_2, Cl_3) to T_G yields the cycle shown in Figure 4.5(b). Note that from Definition 4.38 we have $V(Cl_2) \cap V(Cl_3) \neq \emptyset$. From this new edge (Cl_2, Cl_3) we obtain the following consistency equations:

- (3) $\mu_{Cl_2}(c_{V(Cl_2) \cap V(Cl_3)}) = \mu_{Cl_3}(c_{V(Cl_2) \cap V(Cl_3)}),$ for each configuration $c_{V(Cl_2) \cap V(Cl_3)}$ of $V(Cl_2) \cap V(Cl_3)$.

From Definition 4.41 we have that $V(Cl_2) \cap V(Cl_3) \subseteq V(Cl_1)$. It follows that $V(Cl_2) \cap V(Cl_3) \subseteq V(Cl_1) \cap V(Cl_2)$ and furthermore that $V(Cl_2) \cap V(Cl_3) \subseteq V(Cl_1) \cap V(Cl_3)$. From the consistency equations (1) we get

$$\mu_{Cl_1}(c_{V(Cl_2) \cap V(Cl_3)}) = \mu_{Cl_2}(c_{V(Cl_2) \cap V(Cl_3)}), \quad \text{for each configuration } c_{V(Cl_2) \cap V(Cl_3)} \text{ of } V(Cl_2) \cap V(Cl_3),$$

by further marginalization. From (2) we get

$$\mu_{Cl_1}(c_{V(Cl_2) \cap V(Cl_3)}) = \mu_{Cl_3}(c_{V(Cl_2) \cap V(Cl_3)}), \quad \text{for each configuration } c_{V(Cl_2) \cap V(Cl_3)} \text{ of } V(Cl_2) \cap V(Cl_3).$$

From the transitivity of the equality we have

$$\mu_{Cl_2}(c_{V(Cl_2) \cap V(Cl_3)}) = \mu_{Cl_3}(c_{V(Cl_2) \cap V(Cl_3)}), \quad \text{for each configuration } c_{V(Cl_2) \cap V(Cl_3)} \text{ of } V(Cl_2) \cap V(Cl_3).$$

So, the consistency equations (3) follow immediately from the consistency equations (1) and (2). ■

Analogous to our observations from the preceding section, we have the following relation between global extensions of a set M of partial specifications of marginal distributions and solutions to the joint system of constraints $Dx = m, Tx = 0, x \geq 0$, obtained from a partially quantified belief network $B = (G, M)$.

PROPOSITION 4.46. *Let $B = (G, M)$ be a partially quantified belief network where G is a decomposable graph with the clique set $Cl = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, and where $M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ is a set of partial specifications of marginal distributions. Let $Dx = m$, $Tx = 0$, $x \geq 0$, be the joint system of constraints obtained from B as described in the foregoing. Then, the following properties hold:*

- (1) *For each global extension $M = \{\mu_{Cl_i} \mid Cl_i \in Cl(G)\}$ of M , we have that the vector $x = (x_1, \dots, x_m)$ of subvectors x_i of constituent probabilities of the marginal distribution μ_{Cl_i} associated with clique Cl_i , $i = 1, \dots, m$, is a solution to $Dx = m$, $Tx = 0$, $x \geq 0$.*
- (2) *For each solution $x = (x_1, \dots, x_m)$ to $Dx = m$, $Tx = 0$, $x \geq 0$, we have that each subvector x_i defines a marginal distribution μ_{Cl_i} associated with clique Cl_i , $i = 1, \dots, m$, such that $M = \{\mu_{Cl_i} \mid Cl_i \in Cl(G)\}$ is a global extension of M .*

We recall that in the preceding section we defined a best lower bound function and a best upper bound function relative to a partial specification of a joint probability distribution. We now define similar functions relative to partial specifications of marginal distributions.

DEFINITION 4.47. *Let G be a decomposable graph with the vertex set $V(G) = \{V_1, \dots, V_n\}$, $n \geq 1$, and the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$. Let $\mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_i \mid V_i \in V(G)\}$, and for each clique $Cl_i \in Cl(G)$, let $\mathcal{B}(Cl_i) \subseteq \mathcal{B}(v_1, \dots, v_n)$ be the free Boolean algebra generated by $\{v_j \mid V_j \in V(Cl_i)\}$. Let $M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ be a set of partial specifications of marginal distributions such that $B = (G, M)$ is a partially quantified belief network. For each $m_{Cl_i} \in M$, the function $bub_{Cl_i}: \mathcal{B}(Cl_i) \rightarrow [0, 1]$ defined by $bub_{Cl_i}(c) = \sup \{\mu_{Cl_i}(c) \mid \mu_{Cl_i} \in M \text{ where } M \text{ is a global extension of } M\}$ for all $c \in \mathcal{B}(Cl_i)$, is called the best upper bound function relative to m_{Cl_i} . For each $m_{Cl_i} \in M$, the best lower bound function relative to m_{Cl_i} , denoted by blb_{Cl_i} , is defined symmetrically.*

It will be evident that for each $c \in \mathcal{B}(Cl_i)$ we can find an extension M of M such that $\mu_{Cl_i} \in M$ and $\mu_{Cl_i}(c) = bub_{Cl_i}(c)$; a similar observation can be made concerning blb_{Cl_i} .

In the following discussion it is assumed that a probability of interest $Pr(c)$ is local to a clique, that is, involves variables all occurring in one and the same clique. It will be evident from the foregoing that we have that the problem of

finding the best upper bound for a probability of interest $Pr(c)$ is equivalent to the following linear programming problem:

maximize $Pr(c)$

subject to

$$(i) \quad Dx = m,$$

$$(ii) \quad Tx = 0, \text{ and}$$

$$(iii) \quad x \geq 0.$$

We can solve this linear programming problem using a traditional LP-program. In such a straightforward approach, however, the modular structure of the problem at hand is not exploited. For solving linear programming problems of special structure, several methods have been introduced, see for example [LASD70]. For a problem having the form shown in Figure 4.2, a method for taking advantage of this angular structure has been designed which is known as *Dantzig-Wolfe decomposition*, [PAPA82,LASD70]. This decomposition method basically amounts to solving the entire problem by iteratively solving the separate 'blocks' of the problem; this iteration process is monitored globally by a *master problem*, the major part of which is formed by so-called *coupling constraints* (in our case the consistency equations). If we were to apply Dantzig-Wolfe decomposition to our problem, however, the computations to be performed would not be restricted to local computations per clique only.

Now, recall that our problem is an even special case of the angular form shown in Figure 4.2: each consistency equation involves variables from precisely two cliques only. This observation together with the properties of a clique tree allow us to devise a new decomposition algorithm in which all computations are local to the cliques (provided of course that the probability of interest is local to a clique as well). This decomposition algorithm will be stated in Algorithm 4.48; its correctness will be shown in the Lemmas 4.49 and 4.50. First, however, we describe the basic idea of the algorithm informally for our running example.

Recall that the tree shown in Figure 4.4 is a clique tree of our example graph G . In Figure 4.6 this clique tree is depicted once more, this time explicitly showing the nonempty clique intersections by means of boxes; in this convention we follow Jensen et al. [JENS88b]. Note that this figure also depicts the structure of our linear programming problem: each ellipse may be viewed as representing a system of linear constraints $D_i x_i = m_i$, $x_i \geq 0$, and each box may be viewed as representing a system of consistency equations $T_{i,j} x_i - T_{j,i} x_j = 0$. We now use an object-oriented style of discussion: we view the vertices of the clique tree as *autonomous objects* holding the local systems of constraints $D_i x_i = m_i$, $x_i \geq 0$, as *private data*. These objects are only able to communicate with their direct neighbours in the clique tree and only 'through' the consistency equations: these equations are used for the translation of variables of one clique in terms of variables of another one. So, the edges of the clique tree are viewed as *communication channels*.

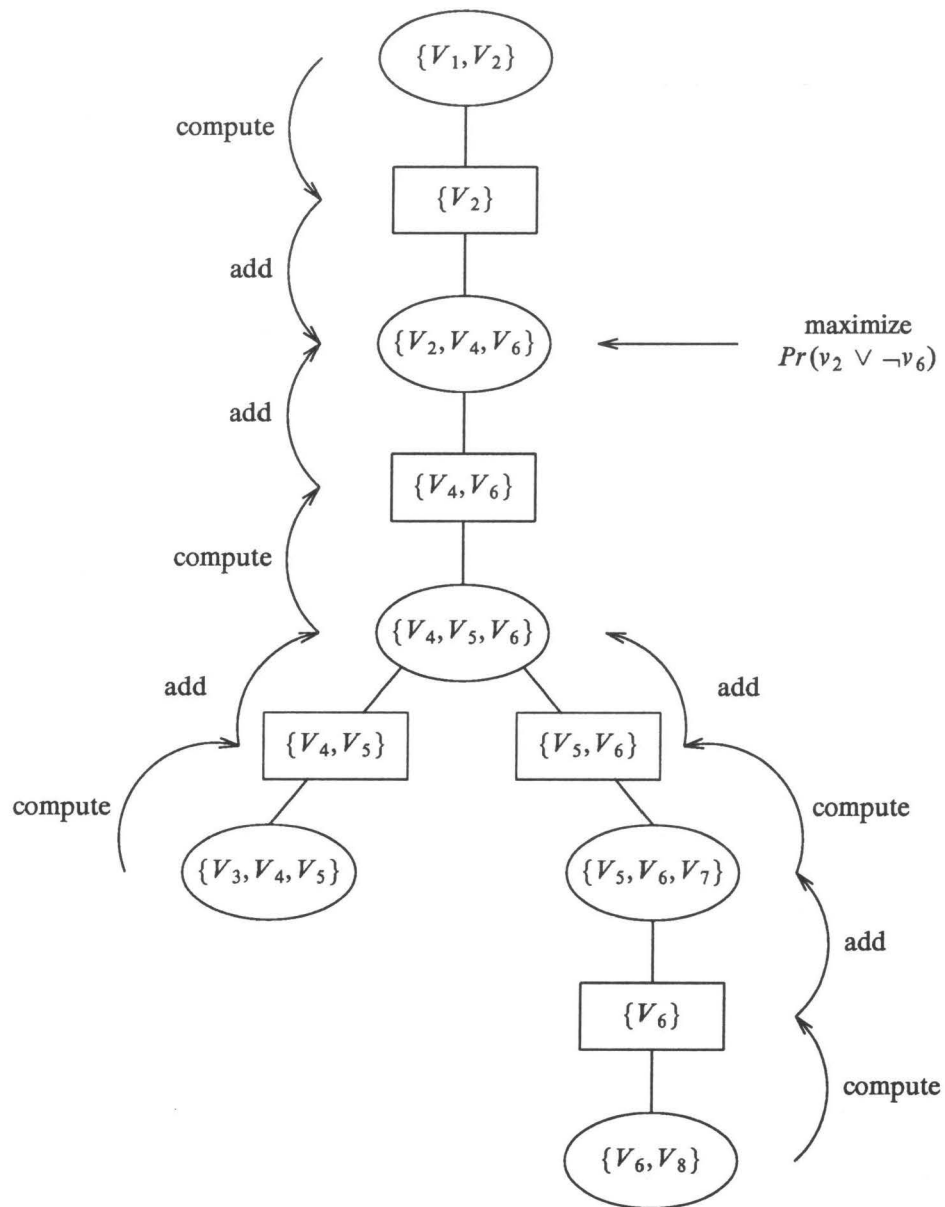


FIGURE 4.6. A decomposition method with local computations.

Suppose we are interested in the best upper bound for a probability which is local to a specific clique, like the one shown in the figure. The object corresponding with the clique now sends a request for information about further constraints, if any, to its neighbours and then waits until it has received the requested information from all of them. For the moment, each 'interior' object in the clique tree just passes the request on to its other neighbours and then awaits the requested information. As soon as a leaf (or the root) of the tree receives such a request for information, a second pass through the tree is started. The leaf computes the feasible set of its local system of linear constraints and derives from it (by means of projection) the set of feasible values for the probabilities which are the constituent probabilities for the intersection with its neighbour. This information then is passed on to this neighbour via the appropriate communication channels using the consistency equations for 'translation' of the variables. This results in the addition of extra constraints to the local system of constraints of this neighbour. These computations are performed by the interior vertices as well until the object that started the computation has been reached again. The arcs in Figure 4.6 represent the flow of computation from this second pass through the clique tree. From its (extended) local system of linear constraints, the object that started the computation may now compute the best upper bound for our probability of interest. We will show that the result thus obtained is the same as when obtained directly from the joint system of constraints. The intuition of this property is that when the process has again reached the object that started the computation, this object has been 'informed' of all constraints of the entire joint system. By directing the same process once more towards the root and the leaves of the tree, all objects can be brought into this state. So, in three passes through the clique tree, each object locally has a kind of global knowledge concerning the entire joint system of constraints. For any probability of interest that is local to a clique we can now compute a probability interval locally.

The following algorithm describes these three passes.

ALGORITHM 4.48. *Let $B = (G, M)$ be a partially quantified belief network as defined in Definition 4.31, where G is a decomposable graph with the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m > 1$, and where $M = \{m_{Cl_i} \mid Cl_i \in Cl(G)\}$ is a set of partial specifications of marginal distributions. Let $D_i x_i = m_i$, $x_i \geq 0$, be the system of linear constraints obtained from m_{Cl_i} , $i = 1, \dots, m$. Let $T_G = (V(T_G), E(T_G))$ be a clique tree of G as defined in Definition 4.41. Let $T_{i,j} x_i - T_{j,i} x_j = 0$ be the system of consistency equations obtained from the edge $(Cl_i, Cl_j) \in E(T_G)$. Now view T_G as a computational architecture in which the vertices are objects and the edges are communication channels. Without loss of generality we assume that the computation is started by the root Cl_s of the clique tree T_G .*

The root Cl_s of T_G starts the computation by performing the following actions:

1. Send a request for information to all neighbours and wait.
2. If a return message, having the form of a system of constraints, has been received from all neighbours, then add these systems of constraints to the local system of constraints $D_s \mathbf{x}_s = \mathbf{m}_s$, $\mathbf{x}_s \geq \mathbf{0}$; compute the feasible set F_s of the resulting system and derive from it the set $\{T_{s,j} \mathbf{x}_s | \mathbf{x}_s \in F_s\}$, for each $Cl_j \in \nu(Cl_s)$.
3. For each such neighbouring clique Cl_j , send this information as a system of constraints to Cl_j using $T_{s,j} \mathbf{x}_s - T_{j,s} \mathbf{x}_j = \mathbf{0}$.

Each leaf Cl_i of T_G performs the following actions:

1. Wait for a message.
2. If a request for information is received, then compute the feasible set F_i of the local system of constraints $D_i \mathbf{x}_i = \mathbf{m}_i$, $\mathbf{x}_i \geq \mathbf{0}$, and derive from it the set $\{T_{i,j} \mathbf{x}_i | \mathbf{x}_i \in F_i\}$, for $Cl_j \in \nu(Cl_i)$.
3. Send this information as a system of constraints to Cl_j using $T_{i,j} \mathbf{x}_i - T_{j,i} \mathbf{x}_j = \mathbf{0}$, and then wait for a message.
4. If a system of linear constraints is received, then add this system to the local system of constraints $D_i \mathbf{x}_i = \mathbf{m}_i$, $\mathbf{x}_i \geq \mathbf{0}$.

For each interior vertex Cl_i , let the vertex Cl_i^+ be defined as $Cl_i^+ \in \nu(Cl_i)$ and Cl_i^+ is on the path from Cl_i to Cl_s , and let the set Cl_i^- be defined as $Cl_i^- = \nu(Cl_i) \setminus \{Cl_i^+\}$. Each interior vertex Cl_i performs the following actions:

1. Wait for a message.
2. If a request for information is received from the neighbour Cl_i^+ , then pass this message on to all other neighbours $Cl_j \in Cl_i^-$.
3. If systems of constraints have been received from all neighbours $Cl_k \in Cl_i^-$ (or from Cl_i^+ , respectively), then add these additional systems of constraints to $D_i \mathbf{x}_i = \mathbf{m}_i$, $\mathbf{x}_i \geq \mathbf{0}$; compute the feasible set F_i of the resulting system of constraints and derive from it the set $\{T_{i,j} \mathbf{x}_i | \mathbf{x}_i \in F_i\}$ for $Cl_j = Cl_i^+$ (or for each $Cl_j \in Cl_i^-$, respectively).
4. For each such clique Cl_j , send this information as a system of constraints to Cl_j using $T_{i,j} \mathbf{x}_i - T_{j,i} \mathbf{x}_j = \mathbf{0}$.

In the following lemma, it is shown that after applying Algorithm 4.48 an equilibrium has been reached.

LEMMA 4.49. Let $B = (G, M)$ be a partially quantified belief network as in the preceding algorithm. For each clique $Cl_i \in Cl(G)$, let $D_i x_i = m_i$, $x_i \geq 0$, be the system of linear constraints obtained from $m_{Cl_i} \in M$, $i = 1, \dots, m$, $m > 1$. Let $T_G = (V(T_G), E(T_G))$ be a clique tree of G as defined in Definition 4.41. Let $T_{i,j} x_i - T_{j,i} x_j = 0$ be the system of equations obtained from the edge $(Cl_i, Cl_j) \in E(T_G)$. Now consider the extended local systems of constraints after applying Algorithm 4.48; for each clique $Cl_i \in Cl(G)$, let F_i^{local} be the feasible set of the extended local system of constraints. Then, for each pair of distinct cliques $Cl_i, Cl_j \in Cl(G)$ such that $V(Cl_i) \cap V(Cl_j) \neq \emptyset$ we have $\{T_{i,j} x_i | x_i \in F_i^{local}\} = \{T_{j,i} x_j | x_j \in F_j^{local}\}$.

PROOF. We prove the lemma by induction on the construction of the clique tree as described in Lemma 4.44.

Induction Basis

Consider the case in which we have a decomposable graph G with two cliques Cl_1 and Cl_2 . Let T_G be a clique tree of G ; T_G has two vertices Cl_1 and Cl_2 , as in the following figure (once more the clique intersection has been shown by means of a box):

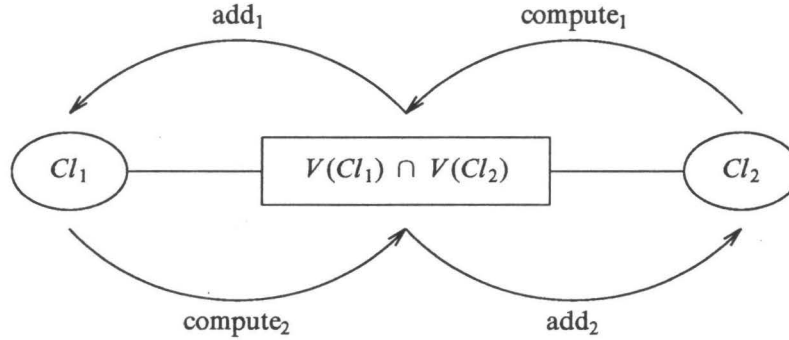


FIGURE 4.7. A clique tree of order 2.

For $i = 1, 2$, let $D_i x_i = m_i$, $x_i \geq 0$, be the local system of linear constraints obtained from the partial specification $m_{Cl_i} \in M$ associated with clique Cl_i ; let F_i be the feasible set of $D_i x_i = m_i$, $x_i \geq 0$. Recall that each vector $x_i \in F_i$ defines a marginal distribution μ_{Cl_i} which is an extension of m_{Cl_i} . Let $T_{1,2} x_1 - T_{2,1} x_2 = 0$ be the system of consistency equations obtained from T_G .

Without loss of generality, we assume that Algorithm 4.48 is applied starting with Cl_1 . Cl_1 sends a request for information to Cl_2 . Cl_2 now computes the feasible set F_2 of its initial local system of constraints $D_2 x_2 = m_2$, $x_2 \geq 0$. It subsequently computes the set $\{T_{2,1} x_2 | x_2 \in F_2\}$, that is, it computes the feasible values for the constituent probabilities for the intersection $V(Cl_1) \cap V(Cl_2)$ obtained from marginalization of a yet unknown marginal distribution μ_{Cl_2} being an extension of m_{Cl_2} . Note that the set

$\{T_{2,1}x_2 \mid x_2 \in F_2\}$ is convex. This information is sent to Cl_1 as a system of constraints; the consistency equations $T_{1,2}x_1 - T_{2,1}x_2 = \mathbf{0}$ are used for the appropriate translation of the variables; Cl_1 adds these constraints to its initial local system of constraints $D_1x_1 = m_1, x_1 \geq \mathbf{0}$. The object Cl_1 then computes the feasible set F_1^{local} of this extended system of constraints. We obviously have that $F_1^{local} \subseteq F_1$. From this feasible set F_1^{local} , the object Cl_1 subsequently derives the set $\{T_{1,2}x_1 \mid x_1 \in F_1^{local}\}$. Note that we have that $\{T_{1,2}x_1 \mid x_1 \in F_1^{local}\} \subseteq \{T_{2,1}x_2 \mid x_2 \in F_2\}$. The object Cl_1 sends this information as a system of (possibly tighter) constraints to Cl_2 , again using $T_{1,2}x_1 - T_{2,1}x_2 = \mathbf{0}$ for translation of the variables. Cl_2 adds these constraints to its local system of constraints and here Algorithm 4.48 halts.

Now suppose that the object Cl_2 actually computes the feasible set F_2^{local} of its extended system of constraints and that it derives from this feasible set the set $\{T_{2,1}x_2 \mid x_2 \in F_2^{local}\}$. It will be evident that we have that $\{T_{2,1}x_2 \mid x_2 \in F_2^{local}\} \subseteq \{T_{1,2}x_1 \mid x_1 \in F_1^{local}\}$. Recall that we have to show that $\{T_{2,1}x_2 \mid x_2 \in F_2^{local}\} = \{T_{1,2}x_1 \mid x_1 \in F_1^{local}\}$. We do so by contradiction. Suppose that there exists a vector x such that $x \in F_1^{local}$ and $x \notin F_2^{local}$. Then, the extended system of constraints of Cl_2 comprises at least one constraint that is not met by x . From $x \in F_1^{local}$ we have that this constraint cannot have been added to the original system of constraints of Cl_2 on account of information received from Cl_1 . So, the violated constraint has to be among the initially specified ones. But then we have that $x \notin F_2$; it follows that $x \notin F_1^{local}$. From the contradiction we conclude that $\{T_{2,1}x_2 \mid x_2 \in F_2^{local}\} = \{T_{1,2}x_1 \mid x_1 \in F_1^{local}\}$. We say that an *equilibrium* has been reached.

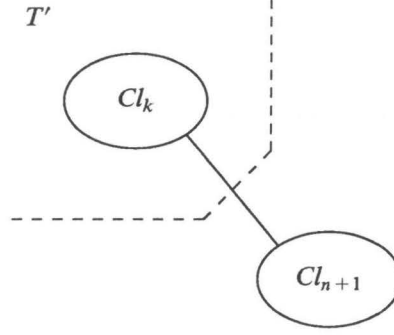
Note that the additional constraints from Cl_1 added to the local system of constraints of Cl_2 might not only affect the bounds on $\mu_{Cl_2}(c_{V(Cl_1) \cap V(Cl_2)})$, but the bounds on $\mu_{Cl_2}(c)$ for any configuration c of a subset of $V(Cl_2) \setminus V(Cl_1)$ as well. But then, since the variables from $V(Cl_2) \setminus V(Cl_1)$ do not occur elsewhere in the clique tree T_G , this cannot possibly lead to new, again tighter constraints on $\mu_{Cl_2}(c_{V(Cl_1) \cap V(Cl_2)})$. So, the computational process cannot 'bounce' from the leaves of the tree.

Induction Hypothesis

Suppose that for some $n \geq 2$, after applying Algorithm 4.48 to a clique tree T_G of a decomposable graph G of order n , we have that for each pair of distinct cliques $Cl_i, Cl_j \in Cl(G)$ such that $V(Cl_i) \cap V(Cl_j) \neq \emptyset$ the property $\{T_{i,j}x_i \mid x_i \in F_i^{local}\} = \{T_{j,i}x_j \mid x_j \in F_j^{local}\}$ holds.

Induction Step

Let T_G be a clique tree of a decomposable graph G of order $n + 1$. Lemma 4.44 allows us to view T_G as constructed by adding a leaf Cl_{n+1} to a clique tree T' of order n . Assume that Cl_{n+1} is connected to a vertex Cl_k of T' as shown in Figure 4.8. Now, let Cl_{n+1} compute the feasible set F_{n+1} of its initial local system of constraints $D_{n+1}x_{n+1} = m_{n+1}, x_{n+1} \geq \mathbf{0}$. It

FIGURE 4.8. A clique tree of order $n + 1$.

subsequently computes the set $\{T_{n+1,k}x_{n+1} \mid x_{n+1} \in F_{n+1}\}$, that is, it computes the feasible values for the constituent probabilities for the intersection $V(Cl_{n+1}) \cap V(Cl_k)$ obtained from marginalization of a yet unknown marginal distribution $\mu_{Cl_{n+1}}$ which is an extension of the partial specification $m_{Cl_{n+1}}$ associated with Cl_{n+1} . The object Cl_{n+1} now sends this information as a system of constraints to Cl_k , using $T_{n+1,k}x_{n+1} - T_{k,n+1}x_k = \mathbf{0}$ for the appropriate translation of the variables. The object Cl_k adds these constraints to its initial local system of constraints.

We view this as an initialization before applying Algorithm 4.48. We now take for Cl_k the thus extended system of constraints and apply the algorithm to T' (disregarding Cl_{n+1}). From the induction hypothesis it follows that after the execution of the algorithm, for each pair of distinct cliques $Cl_i, Cl_j \in Cl(G) \setminus \{Cl_{n+1}\}$ such that $V(Cl_i) \cap V(Cl_j) \neq \emptyset$, we have that $\{T_{i,j}x_i \mid x_i \in F_i^{local}\} = \{T_{j,i}x_j \mid x_j \in F_j^{local}\}$.

Now let Cl_k compute the feasible set F_k^{local} of its (extended) local system of constraints after application of Algorithm 4.48 to T' . From this set F_k^{local} , the object Cl_k subsequently computes the set $\{T_{k,n+1}x_k \mid x_k \in F_k^{local}\}$. The object Cl_k now sends these (possibly tighter) constraints to Cl_{n+1} . After using $T_{n+1,k}x_{n+1} - T_{k,n+1}x_k = \mathbf{0}$ for translation of the variables, Cl_{n+1} adds these constraints to its local system of constraints. Since Cl_k was extended initially with the additional constraints from Cl_{n+1} before applying Algorithm 4.48 to T' we have $\{T_{k,n+1}x_k \mid x_k \in F_k^{local}\} \subseteq \{T_{n+1,k}x_{n+1} \mid x_{n+1} \in F_{n+1}\}$. Now suppose that Cl_{n+1} actually computes the feasible set F_{n+1}^{local} of its thus extended system of constraints. Using the same argument as in the induction basis, we have that the computation process cannot ‘bounce’ from Cl_{n+1} ; so, $\{T_{k,n+1}x_k \mid x_k \in F_k^{local}\} = \{T_{n+1,k}x_{n+1} \mid x_{n+1} \in F_{n+1}^{local}\}$. It now follows from further marginalization and the transitivity of the equality that for each pair of distinct cliques $Cl_i, Cl_j \in Cl(G)$ such that $V(Cl_i) \cap V(Cl_j) \neq \emptyset$, we have that $\{T_{i,j}x_i \mid x_i \in F_i^{local}\} = \{T_{j,i}x_j \mid x_j \in F_j^{local}\}$.

Recall that in this induction step we added the constraints obtained from Cl_{n+1} to Cl_k before applying Algorithm 4.48 to the clique tree T' of order n . However, if we apply the algorithm to the entire clique tree T starting with the root of T' , then the object Cl_k has to wait for systems of constraints from *all* its successors before it can compute from its extended system of constraints, the constraints to be sent to its predecessor in the clique tree. So, in both cases Cl_k computes the constraints to be sent to its predecessor from the same system of constraints. We therefore have that an equilibrium is reached after applying Algorithm 4.48 to the clique tree T of order $n + 1$. ■

In the following lemma it is shown that the decomposition algorithm yields the correct results.

LEMMA 4.50. *Let $B = (G, M)$ be a partially quantified belief network as in Algorithm 4.48. Let $D_i x_i = m_i$, $x_i \geq 0$, be the system of linear constraints obtained from $m_{Cl_i} \in M$, $Cl_i \in Cl(G)$, $i = 1, \dots, m$, $m > 1$. Let $T_G = (V(T_G), E(T_G))$ be a clique tree of G . Let $Tx = 0$ be the system of consistency equations obtained from T_G and let $Dx = m$, $Tx = 0$, $x \geq 0$, be the joint system of constraints, having the feasible set F^{joint} . For each Cl_i , let $F_i^{joint} = \{x_i \mid (x_1, \dots, x_i, \dots, x_m) \in F^{joint}\}$, $i = 1, \dots, m$. Furthermore, let F_i^{local} be the feasible set of the extended system of constraints for clique Cl_i after application of Algorithm 4.48. Then, for $i = 1, \dots, m$, we have $F_i^{local} = F_i^{joint}$.*

PROOF. For $i = 1, \dots, m$, let F_i be the feasible set of the local system of linear constraints $D_i x_i = m_i$, $x_i \geq 0$, before application of Algorithm 4.48. From the proof of the preceding lemma it will be evident that we have $F_i^{local} \subseteq F_i$, $i = 1, \dots, m$. From the observation that each $D_i x_i = m_i$, $x_i \geq 0$, $i = 1, \dots, m$, is a subsystem of the joint system of constraints $Dx = m$, $Tx = 0$, $x \geq 0$, we also have $F_i^{joint} \subseteq F_i$. We will show that $F_i^{local} \subseteq F_i^{joint}$ and $F_i^{joint} \subseteq F_i^{local}$.

- (1) Consider a vector $x_i' \in F_i^{local}$. Recall that x_i' defines a marginal distribution μ'_{Cl_i} which is an extension of the partial specification m_{Cl_i} associated with Cl_i . From the equilibrium property proven in the preceding lemma, we have that we can find a vector $x_j' \in F_j^{local}$ such that $T_{i,j} x_i' - T_{j,i} x_j' = 0$ using the consistency equations obtained from the edge (Cl_i, Cl_j) in $E(T_G)$. Recursively repeating the argument, we have that there exists a vector $x' = (x_1', \dots, x_m')$, $x_j' \in F_j^{local}$, $j = 1, \dots, m$, such that $Tx' = 0$, $x' \geq 0$. From $F_i^{local} \subseteq F_i$, $i = 1, \dots, m$, it furthermore follows that x' is a solution to $Dx = m$, $x \geq 0$. So, x' is a solution to $Dx = m$, $Tx = 0$, $x \geq 0$. But then, we have that $x_i' \in F_i^{joint}$. It follows that $F_i^{local} \subseteq F_i^{joint}$.
- (2) Consider a vector $x_i' \in F_i^{joint}$. We have that there exists at least one vector $x' = (x_1', \dots, x_i', \dots, x_m')$ which is a solution to the joint system of constraints $Dx = m$, $Tx = 0$, $x \geq 0$. We have for each $k = 1, \dots, m$, that x_k' is a solution to $D_k x_k = m_k$, $x_k \geq 0$; so,

$x_k' \in F_k$. Now consider a subvector x_j' of x' corresponding with a clique Cl_j which is a neighbour of Cl_i in T_G . Since x' is a solution to $Tx = 0$, we have that $T_{i,j}x_i' - T_{j,i}x_j' = 0$. Now suppose that after applying Algorithm 4.48, $x_i' \notin F_i^{local}$. From the equilibrium property proven in the previous lemma, it follows that $x_j' \notin F_j^{local}$. Recursively repeating the argument, we have for $k = 1, \dots, m$, that $x_k' \notin F_k^{local}$. But then, for some j this contradicts $x_j' \in F_j^{joint}$. It follows that $x_i' \in F_i^{local}$. We have $F_i^{joint} \subseteq F_i^{local}$.

■

Note that after Algorithm 4.48 has been applied, for any local probability a lower bound and an upper bound can be computed locally from an appropriate (extended) local system of constraints. The resulting probability interval may be rather wide, in fact it may be too wide for practical purposes. However, it is in a sense an 'honest' result: it just reflects the lack of knowledge concerning the joint probability distribution. Note that if we had selected a single solution to the joint system of constraints for computing probabilities of interest, the resulting 'point' probabilities would have given the false impression of complete information.

Now recall that in this section we have taken a partially quantified belief network to consist of a decomposable graph G and a number of probabilities which are local to the cliques of G . Viewed in the context of such a partially quantified belief network, the probability intervals resulting from optimizing local probabilities after Algorithm 4.48 has been applied are exact. The decomposable graph G , however, has been obtained from an acyclic directed graph by applying the transformation scheme proposed by Lauritzen and Spiegelhalter. Now recall from Section 3.4 that the decomposable graph does not reflect all independency relationships between the statistical variables discerned that have been assessed initially by the domain expert: it may contain several 'dummy' edges. It will be evident that in this case the probability intervals may be too pessimistic, since we have not taken all independency relationships into account. For obtaining exact probability intervals, however, the entire approach can be applied recursively to the cliques containing such 'dummy' edges.

We conclude this section with a discussion of some issues concerning the computational complexity of the presented algorithm. In this discussion, n denotes the number of statistical variables in the partially quantified belief network $B = (G, M)$, that is, n is the order of the decomposable graph G . Algorithm 4.48 essentially is composed of a sequence of solving smaller problems. The computational complexity of the algorithm therefore is dependent upon the number of problems to be solved as well as upon the complexity of solving these separate problems. It will be evident that in general the algorithm may take exponential time. However, if the maximal clique size is small compared to the number of statistical variables, that is, if the maximal clique size is bound by some constant k , then the algorithm will take polynomial time. The fact is, that under this restriction on the clique

sizes, computing the required information from a local system of constraints will only take constant time. Moreover, the system has to solve $O(n)$ problems of constant size, since the maximum number of cliques is n . Note that after Algorithm 4.48 has been applied, any local probability can be optimized locally; under the restriction mentioned above this takes constant time. It is noted that in [LAUR88a], Lauritzen and Spiegelhalter mention the same restriction on the maximal clique size for their method to be feasible. An important question for the decomposition algorithm to be of practical use is the question whether it is likely that the mentioned restriction will be met in practice. Concerning this question, Pearl argues that sparse, irregular networks are generally appropriate in practical situations, [PEAR88].

4.4. PROCESSING EVIDENCE

Recall from the introduction to this chapter that in order to be able to exploit a partially quantified belief network for reasoning with uncertainty, we had to devise a method for deriving information about probabilities of interest from the network and a method for processing evidence. In this section we address the latter problem of propagating evidence through a partially quantified belief network. Again we assume that the graphical part of such a belief network has been transformed into a decomposable graph. In addition, we assume that a clique tree of the decomposable graph has been constructed. We will once more use an object-oriented style of discussion and view the clique tree as a computational architecture just like we have done in the preceding section. We assume that each of the autonomous objects of the clique tree holds a local system of linear constraints that initially has been obtained from the appropriate partial specification of a marginal distribution and subsequently has been extended using Algorithm 4.48.

Now suppose that a piece of evidence becomes available. We discern two types of evidence:

- (1) evidence concerning a certain partial specification of a marginal distribution, called *case-independent* evidence, and
- (2) evidence observed for a specific case, called *case-dependent* evidence.

Case-independent evidence is merely *new* knowledge concerning a certain partial specification of a marginal distribution rendering it 'more specified': it is information we did not have before. This type of evidence is dealt with just by adding another constraint representing the piece of evidence to the appropriate system of constraints obtained from that specific partial specification. After application of Algorithm 4.48 new bounds on any local probability of interest can be computed locally. The bounds obtained after processing this type of evidence are modified monotonically: new evidence merely leads to the same or narrower probability intervals. Note that this property allows for stepwise filling-in a quantification of a belief network.

The second type of evidence concerns information that for the specific case we are looking at we have observed that a certain statistical variable has a

certain value. From the way an updated probability distribution in the fully specified case is computed, it will be evident that we cannot simply add this evidence as a new constraint to an appropriate system of constraints; the addition of such a constraint will probably even render the system infeasible.

Now, recall from the previous section that for computing bounds on probabilities of interest from a partially quantified belief network we exploited the work on belief networks by S.L. Lauritzen and D.J. Spiegelhalter. Once more we take their model as a point of departure. Adhering to the basic idea of the scheme for evidence propagation used in this model, our aim now is to arrive at a method for 'updating' the separate systems of constraints locally, yielding new systems of linear constraints such that each of these systems defines the possible extensions of the corresponding partial specification of a marginal distribution after it has been updated with the evidence. Then, after a piece of evidence has been processed we can compute bounds on probabilities of interest locally just like before the evidence was processed: the notion of a partially quantified belief network is invariant under evidence propagation. Unfortunately, we have not been able to find such a method for processing case-dependent evidence. The problem of evidence propagation through a partially quantified belief network will have to be a subject of future research. Here, we will merely state some of the problems we have encountered when trying to devise such a propagation method. In Section 4.4.1 we will discuss processing a piece of case-dependent evidence in one clique of the graph; in Section 4.4.2 we will consider the propagation of the evidence to the other cliques of the graph and show by means of an easy counterexample that the scheme for evidence propagation proposed by Lauritzen and Spiegelhalter cannot be extended to deal with probability intervals as indicated.

4.4.1. Processing Evidence in One Clique

We consider a partially quantified belief network $B = (G, M)$ as defined in the previous section. Let $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m \geq 1$, be the clique set of the decomposable graph G of B . For each clique Cl_i , let $m_{Cl_i} \in M$ be the partial specification of a marginal distribution associated with Cl_i . Now, for each Cl_i , we obtain a system of linear constraints from the appropriate partial specification m_{Cl_i} ; recall that each solution vector of such a local system of constraints defines a marginal distribution that is an extension of m_{Cl_i} . Furthermore, let F_{Cl_i} denote the feasible set of the (extended) local system of constraints for Cl_i resulting after applying Algorithm 4.48, $i = 1, \dots, m$. Now suppose that we obtain the case-dependent evidence that the statistical variable $V \in V(G)$ has the value *true* (the case in which we have observed that V has the value *false* is dealt with analogously). Let Cl be a clique in G containing V and let r be the number of statistical variables in the vertex set of Cl , $r \geq 1$.

We consider the marginal distribution μ_{Cl} defined by a specific vector $x \in F_{Cl}$ and investigate the updating of μ_{Cl} . Note that 2^{r-1} constituent

(marginal) probabilities of μ_{Cl} specify v and that the remaining 2^{r-1} ones specify $\neg v$. It will be evident that updating the marginal distribution μ_{Cl} amounts to setting all constituent probabilities specifying $\neg v$ equal to zero, and then normalizing the remaining constituent probabilities in order to render the result again a marginal distribution. Now consider the defining vector x . Without loss of generality, we take the components of x to be ordered in such a way that the first 2^{r-1} components correspond to the constituent probabilities of μ_{Cl} specifying v and that the remaining components correspond to the constituent probabilities specifying $\neg v$. The following definition introduces an update mapping such that when applied to the vector x the updated vector defines the updated marginal distribution (for ease of exposition, we take the updated vector to be of the same dimension as the original one).

DEFINITION 4.51. *The update mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, is the partial mapping defined by*

- $$(1) \quad U((x_1, \dots, x_{2^r})) = \left(\frac{x_1}{\sum_{i=1}^{2^{r-1}} x_i}, \dots, \frac{x_{2^{r-1}}}{\sum_{i=1}^{2^{r-1}} x_i}, 0, \dots, 0 \right) \text{ if } \sum_{i=1}^{2^{r-1}} x_i \neq 0, \text{ and}$$
- (2) $U(x) = \text{undefined}$, otherwise.

The case in which we apply the update mapping U to some vector $x \in \mathbb{R}^{2^r}$ of constituent (marginal) probabilities for which we have $\sum_{i=1}^{2^{r-1}} x_i = 0$ deserves some special attention. For the marginal distribution μ_{Cl} defined by such a vector x , we evidently have $\mu_{Cl}(v) = 0$. Evidence that $\mu_{Cl}(v) = 1$ contradicts this prior information. For this case, we take $U(x) = \text{undefined}$; this is an arbitrary choice. We return to this observation shortly.

Since we are primarily interested in applying mappings to vectors representing marginal distributions, we will frequently restrict the discussion to unit simplices.

DEFINITION 4.52. *The unit simplex in \mathbb{R}^n , $n \geq 1$, denoted by S_n , is the convex set in the positive orthant of \mathbb{R}^n such that for each $x \in S_n$ we have $\sum_{i=1}^n x_i = 1$.*

The following lemma states the evident property that when applied to a vector representing a marginal distribution the update mapping U yields a vector which again represents a marginal distribution, provided of course that the result is defined.

LEMMA 4.53. *Let the mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined as above. For each $x \in S_{2^r}$, we have that either $U(x) \in S_{2^r}$ or $U(x) = \text{undefined}$.*

Until now we have only looked at the updating of a single vector. Recall that such a vector is an element of a convex polytope F_{Cl} of vectors, each defining a marginal distribution which is an extension of the initially given partial specification m_{Cl} associated with the clique Cl . For processing the evidence that the statistical variable V has adopted the value *true*, we have to apply the update mapping U to each vector $x \in F_{Cl}$. We therefore are interested in the image $U(F_{Cl})$ of F_{Cl} . Since the mapping U is non-linear, the question arises whether the image of a convex polytope under U again is a convex polytope. We will show that this question may be answered in the affirmative.

It will be evident from the preceding informal discussion that the update mapping U is composed of a multiplication and a projective mapping. The multiplication mapping is defined in the following definition.

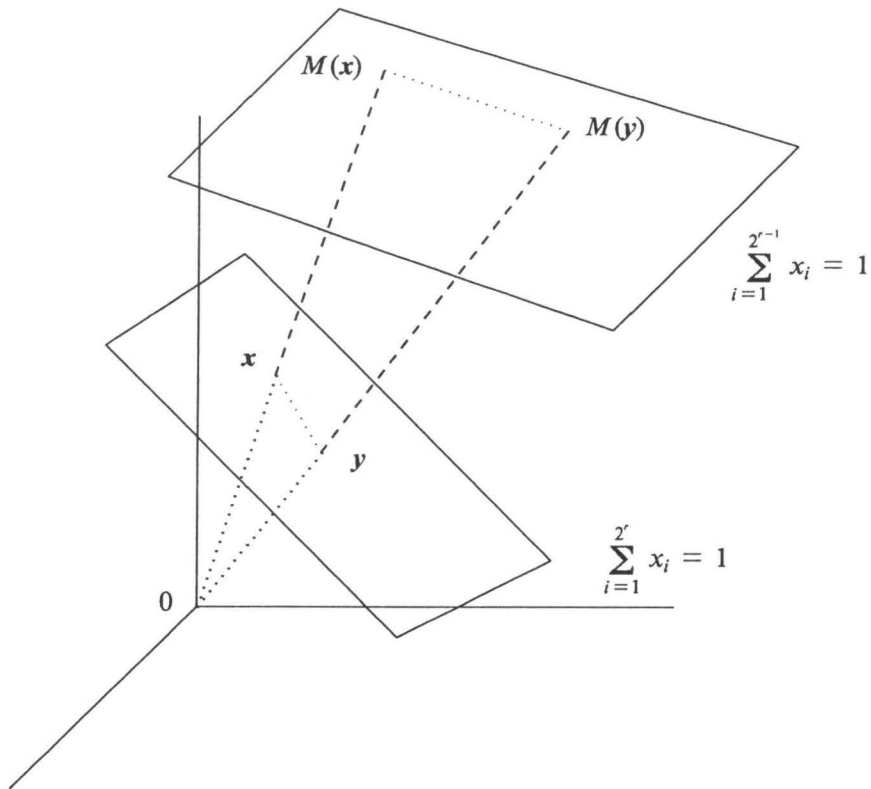


FIGURE 4.9. The general idea of the mapping M .

DEFINITION 4.54. The multiplication mapping $M: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, is the partial mapping defined by

- (1) $M((x_1, \dots, x_{2^r})) = \left[\frac{x_1}{\sum_{i=1}^{2^{r-1}} x_i}, \dots, \frac{x_{2^r}}{\sum_{i=1}^{2^{r-1}} x_i} \right]$ if $\sum_{i=1}^{2^{r-1}} x_i \neq 0$, and
- (2) $M(x) = \text{undefined}$, otherwise.

The geometrical idea of applying the multiplication mapping M to a vector from the unit simplex S_{2^r} is sketched in Figure 4.9. We consider this mapping M in some detail.

LEMMA 4.55. Let the multiplication mapping $M: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.54. Furthermore, let $F \subseteq S_{2^r}$ be a convex polytope such that for each $x \in F$ we have that $M(x) \neq \text{undefined}$. Then, the image $M(F)$ of F is a convex polytope.

PROOF. We prove the lemma by showing that by applying M each line segment in F is mapped again into a line segment.

Let $x, y \in F$. Let $M(x)_i$ denote the i -th component of $M(x)$. Then, $\lambda x + (1 - \lambda)y$, $0 \leq \lambda \leq 1$, represents the line segment between the two points x and y . Consider $M(\lambda x + (1 - \lambda)y)$; note that the conditions of the lemma guarantee that $M(\lambda x + (1 - \lambda)y)$ is defined. We have to show that there exists a scalar μ , $0 \leq \mu \leq 1$, such that the property $M(\lambda x + (1 - \lambda)y) = \mu M(x) + (1 - \mu)M(y)$ holds. We have

$$\begin{aligned} M(\lambda x + (1 - \lambda)y)_i &= \\ &= \frac{\lambda x_i + (1 - \lambda)y_i}{\sum_{j=1}^{2^{r-1}} (\lambda x_j + (1 - \lambda)y_j)} = \frac{\lambda x_i + (1 - \lambda)y_i}{\lambda \sum_{j=1}^{2^{r-1}} x_j + (1 - \lambda) \sum_{j=1}^{2^{r-1}} y_j} \end{aligned}$$

for $i = 1, \dots, 2^r$. Now let $\alpha = \sum_{j=1}^{2^{r-1}} x_j$ and $\beta = \sum_{j=1}^{2^{r-1}} y_j$. Then,

$$\begin{aligned} M(\lambda x + (1 - \lambda)y)_i &= \\ &= \frac{\lambda x_i + (1 - \lambda)y_i}{\lambda \alpha + (1 - \lambda)\beta} = \frac{\lambda}{\lambda \alpha + (1 - \lambda)\beta} x_i + \frac{(1 - \lambda)}{\lambda \alpha + (1 - \lambda)\beta} y_i \end{aligned}$$

By definition, we have $M(x)_i = \frac{x_i}{\alpha}$ and $M(y)_i = \frac{y_i}{\beta}$. So,

$$\begin{aligned} M(\lambda x + (1 - \lambda)y)_i &= \\ &= \frac{\lambda\alpha}{\lambda\alpha + (1 - \lambda)\beta} M(x)_i + \frac{(1 - \lambda)\beta}{\lambda\alpha + (1 - \lambda)\beta} M(y)_i = \\ &= \mu M(x)_i + (1 - \mu)M(y)_i \end{aligned}$$

for $i = 1, \dots, 2^r$, where $\mu = \frac{\lambda\alpha}{\lambda\alpha + (1 - \lambda)\beta}$. Note that $0 \leq \mu \leq 1$. Furthermore, note that $\lambda = 0$ corresponds with $\mu = 0$ and that $\lambda = 1$ corresponds with $\mu = 1$. We have that $\mu M(x) + (1 - \mu)M(y)$ is the line segment between the points $M(x)$ and $M(y)$. It follows that $M(F)$ is convex. ■

LEMMA 4.56. *Let the multiplication mapping $M: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.54. Furthermore, let $F \subseteq S_{2^r}$ be a convex polytope such that for each $x \in F$ we have that $M(x) \neq \text{undefined}$. Then, x is a vertex of F if and only if $M(x)$ is a vertex of $M(F)$.*

PROOF. The lemma follows immediately from the proof of the previous lemma and the observation that for each $x, y \in F$ such that $x \neq y$, we have $M(x) \neq M(y)$. ■

In the previous two lemmas, we have excepted the case in which the multiplication mapping M is applied to some vector $x \in S_{2^r}$ for which $\sum_{i=1}^{2^{r-1}} x_i = 0$. We have pointed out before that this case is somewhat problematic. We now exploit the geometrical view to the mapping M presented informally in Figure 4.9 in order to look at this special case in more detail.

LEMMA 4.57. *Let F denote the feasible set of the system of constraints*

$$\begin{aligned} \sum_{i=1}^{2^r} d_{j,i} x_i &= p_j, \quad j = 1, \dots, k, \quad k \geq 1, \\ \sum_{i=1}^{2^r} x_i &= 1, \text{ and} \\ x_i &\geq 0, \quad i = 1, \dots, 2^r, \end{aligned}$$

where $r \geq 1$. Furthermore, let F° be the feasible set of the system of constraints

$$\sum_{i=1}^{2^r} (d_{j,i} - p_j) x_i = 0, j = 1, \dots, k, k \geq 1,$$

$$\sum_{i=1}^{2^r} x_i = 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r.$$

Then, $F = F^\circ$.

PROOF. The lemma follows from the observation that $\sum_{i=1}^{2^r} d_{j,i} x_i = p_j \cdot \sum_{i=1}^{2^r} x_i$, $j = 1, \dots, k$. ■

Consider the property stated in the preceding lemma once more. From Lemma 4.6 we have that the feasible set C of the system of constraints

$$\sum_{i=1}^{2^r} (d_{j,i} - p_j) x_i = 0, j = 1, \dots, k, k \geq 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r,$$

where $r \geq 1$, is a polyhedral cone. In fact, C equals the polyhedral cone generated by $\text{vert}(F)$, where F is the feasible set of the original system of constraints as in Lemma 4.57. We have that by intersecting this polyhedral cone C with the hyperplane $\sum_{i=1}^{2^r} x_i = 1$, we again obtain this feasible set F .

Now, let M be the multiplication mapping defined according to Definition 4.54. If F does not contain any vectors x for which $M(x)$ is undefined, then we obtain the image $M(F)$ of F by intersecting the cone C with the hyperplane $\sum_{i=1}^{2^r} x_i = 1$. This property is stated more formally in the following lemma.

LEMMA 4.58. Let the multiplication mapping $M: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.54. Let F be the feasible set of the system of constraints

$$\sum_{i=1}^{2^r} d_{j,i} x_i = p_j, j = 1, \dots, k, k \geq 1,$$

$$\sum_{i=1}^{2^r} x_i = 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r,$$

such that for each $x \in F$ we have that $M(x) \neq \text{undefined}$. Let F^* be the feasible set of the system of constraints

$$\sum_{i=1}^{2^r} (d_{j,i} - p_j) x_i = 0, j = 1, \dots, k, k \geq 1,$$

$$\sum_{i=1}^{2^{r-1}} x_i = 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r.$$

Then, $F^* = M(F)$.

PROOF. We will show that $M(F) \subseteq F^*$; the proof that $F^* \subseteq M(F)$ is analogous.

Recall from the conditions of the lemma that we have assumed that $M(x)$ is defined for each $x \in F$. Now, let $x' \in F$. We have that $M(x') = \lambda' x'$ where $\lambda' = \frac{1}{2^{r-1}}$. From $x' \in F$, we have that

$$\sum_{i=1}^{2^r} x_i$$

$$\sum_{i=1}^{2^r} d_{j,i} x'_i = p_j, j = 1, \dots, k, \text{ and}$$

$$\sum_{i=1}^{2^r} x'_i = 1, \text{ and}$$

$$x'_i \geq 0, i = 1, \dots, 2^r.$$

It follows that

$$\sum_{i=1}^{2^r} (d_{j,i} - p_i) x'_i = 0, j = 1, \dots, k.$$

Evidently, we have that

$$\sum_{i=1}^{2^r} (d_{j,i} - p_i)(\lambda x'_i) = 0, j = 1, \dots, k,$$

for any $\lambda \geq 0$. For $\lambda = \lambda'$ we furthermore have

$$\sum_{i=1}^{2^{r-1}} (\lambda' x'_i) = 1.$$

So, $\lambda' x' \in F^*$. ■

We return once more to the case in which the feasible set F of the (original) system of constraints as in Lemma 4.58 comprises at least one solution vector \mathbf{x} for which the image $M(\mathbf{x}) = \text{undefined}$. Now, consider the polyhedral cone C being the feasible set of the system of constraints

$$\sum_{i=1}^{2^r} (d_{j,i} - p_j) x_i = 0, j = 1, \dots, k, k \geq 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r.$$

We have that C comprises at least one solution vector \mathbf{x} for which $\sum_{i=1}^{2^{r-1}} x_i = 0$.

It follows that the feasible set F^* of the system of constraints

$$\sum_{i=1}^{2^r} (d_{j,i} - p_j) x_i = 0, j = 1, \dots, k,$$

$$\sum_{i=1}^{2^{r-1}} x_i = 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r,$$

is either empty or unbounded (note that this property follows from $\sum_{i=1}^{2^{r-1}} x_i = 0$ and $\sum_{i=1}^{2^{r-1}} x_i = 1$ being parallel hyperplanes). The aptness of this geometrical observation may readily be seen by examining the images $M(\mathbf{y})$ of all vectors $\mathbf{y} \in F$ in the ϵ -neighbourhood $N_\epsilon(\mathbf{x})$ of \mathbf{x} .

Now, recall that we are investigating the question whether the image of a convex polytope under the update mapping U from Definition 4.51 again is a convex polytope. We have argued that this mapping U is composed of the multiplication mapping M and a projective mapping. We now turn our attention to this projective mapping.

DEFINITION 4.59. *The projective mapping $P: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, is the mapping defined by*

$$P((x_1, \dots, x_{2^r})) = (x_1, \dots, x_{2^{r-1}}, 0, \dots, 0).$$

Note that if we take $U = P \circ M$, we formally have to deal with the case where $M(\mathbf{x})$ is undefined for some vector \mathbf{x} . For ease of exposition we disregard such cases.

LEMMA 4.60. *Let the projective mapping $P: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined as above. Let $F \subseteq S_{2^r}$ be a convex polytope. Then, the image $P(F)$ of F is a convex polytope.*

PROOF. The lemma follows from the observation that P is a linear mapping. ■

The following lemma should be evident.

LEMMA 4.61. *Let the projective mapping $P: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined as above. Let $F \subseteq S_{2^r}$ be a convex polytope. Then, $P(x)$ is a vertex of $P(F)$ only if x is a vertex of F .*

Note that the reverse property does not hold, that is, not every vertex x of F corresponds with a vertex of $P(F)$.

We look once more at the problematic case in which the feasible set F of the original system of constraints contains at least one solution vector x for which $M(x) = \text{undefined}$. Recall that for this case we have that the feasible set F^* of the system of constraints

$$\sum_{j=1}^{2^r} (d_{j,i} - p_j) x_i = 0, j = 1, \dots, k, k \geq 1,$$

$$\sum_{i=1}^{2^{r-1}} x_i = 1, \text{ and}$$

$$x_i \geq 0, i = 1, \dots, 2^r,$$

is either empty or unbounded. We observe that if F^* is unbounded, then the image $P(F^*)$ of F^* is bounded since the hyperplane $\sum_{i=1}^{2^{r-1}} x_i = 1$ and the hyperplane defined by $x_i = 0, i = 2^{r-1} + 1, \dots, 2^r$, are orthogonal.

We now combine the Lemmas 4.55, 4.56, 4.60 and 4.61 to yield the following lemma concerning U .

LEMMA 4.62. *Let the update mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined as in Definition 4.51. Let $F \subseteq S_{2^r}$ be a convex polytope such that for each $x \in F$ we have that $U(x) \neq \text{undefined}$. Then,*

- (1) *the image $U(F)$ of F is a convex polytope, and*
- (2) *$U(x)$ is a vertex of $U(F)$ only if x is a vertex of F .*

The following lemma states a more detailed result concerning the image of a convex polytope under the update mapping.

LEMMA 4.63. *Let the update mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.51. Let $F \subseteq S_{2^r}$ be a convex polytope such that for each $x \in F$ we have that $U(x) \neq \text{undefined}$. Let $\text{vert}(F)$ be the set of vertices of F . Then, $U(F) = \text{hull}(U(\text{vert}(F)))$.*

PROOF. We have from Lemma 4.62(2) that the set $U(\text{vert}(F))$ contains all vertices of $U(F)$. The lemma now follows from this observation and Lemma 4.4. ■

Consider the statement of the preceding lemma once more. It will be evident that for a given polytope F having the mentioned property, the set $U(\text{vert}(F))$ is not the minimal spanning set of $U(F)$ since it may contain some interior points from $U(F)$ as well.

In the beginning of this section we have argued that for propagating case-dependent evidence through a partially quantified belief network $B = (G, M)$ we aim at devising a method for 'updating' the (extended) local systems of constraints associated with the cliques of G , yielding new systems of constraints such that each of these defines the possible extensions of the corresponding partial specification of a marginal distribution after it has been updated with the evidence. The last lemma now provides us with a (theoretical) means for updating the system of constraints associated with the clique Cl of G in which the evidence has been entered. In the following algorithm we exploit the property stated in this lemma. Note that we once more assume that the feasible set of the system of constraints does not comprise any vectors x for which $U(x)$ is undefined.

ALGORITHM 4.64. *Let the update mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.51. Let $F \subseteq S_{2^r}$ be a convex polytope such that for each $x \in F$ we have that $U(x) \neq \text{undefined}$. Then, the following algorithm yields a system of constraints having $U(F)$ for its feasible set:*

1. *Compute the set $\text{vert}(F)$ of all vertices of F .*
2. *Apply the operator U to each element $x \in \text{vert}(F)$, thus obtaining the set $U(\text{vert}(F))$.*
3. *Use $U(\text{vert}(F))$ to span the convex hull $U(F) = \text{hull}(U(\text{vert}(F)))$.*
4. *Construct the supporting hyperplanes of $U(F)$ and generate the appropriate system of constraints.*

This algorithm for processing case-dependent evidence is rather inefficient; step (1) in itself already takes exponential time. It shows however that updating a local system of constraints can actually be achieved.

We return once more to the case in which the feasible set F_{Cl} of the (extended) system of constraints associated with the clique Cl contains at least one vector x for which the image $U(x)$ is undefined. We distinguish two cases: if $\text{vert}(F_{Cl})$ consists of only vectors for which the image under U is undefined, then the observed evidence we are trying to process evidently is inconsistent

with the prior information; the evidence cannot be processed and the detected inconsistency should be reported. If, on the other hand, $\text{vert}(F_{Cl})$ also comprises some vertices x for which $U(x)$ is defined, then the observed evidence can be processed. In computing $U(F_{Cl})$, however, we have to exclude all vectors defining marginal distributions the piece of evidence is inconsistent with. From the geometrical observations in the foregoing, it will be evident that the above-mentioned algorithm yields the correct result after just ignoring those vertices of F for which the image under U is undefined.

4.4.2. Propagating Evidence Through a Partially Quantified Belief Network

From the introduction to this section we recall that our aim is to arrive at a method for processing evidence in a partially quantified belief network that is similar in concept to the method for evidence propagation presented by Lauritzen and Spiegelhalter for fully quantified networks. In the foregoing we have presented a method for processing a piece of case-dependent evidence in one clique of a given partially quantified belief network; this method basically amounted to ‘updating’ the (extended) local system of constraints associated with this clique. We now like to propagate the piece of evidence through the remainder of the network, adhering to the same basic idea. We have mentioned before that we have not been able to devise such a propagation method. In this subsection we state some problems we have encountered in trying to find such a method and show by means of an easy counterexample that the method of Lauritzen and Spiegelhalter cannot be extended to apply to a partially quantified belief network.

Let $B = (G, M)$ be a partially quantified belief network where G is a decomposable graph with the clique set $Cl(G) = \{Cl_1, \dots, Cl_m\}$, $m > 1$. For each clique Cl_i , we obtain a system of linear constraints from the appropriate partial specification $m_{Cl_i} \in M$ associated with Cl_i in the manner described in Section 4.3 and extend it using Algorithm 4.48. The general idea of the propagation of a piece of case-dependent evidence through the entire network is sketched in Figure 4.10. The upper row of the figure shows three cliques Cl_r , Cl_s and Cl_t as they occur in a given clique tree T_G of the decomposable graph G of the network. From now on, let r denote the number of statistical variables in the vertex set of clique Cl_r , let s be the number of variables in clique Cl_s and t the number of variables in Cl_t , $r, s, t \geq 1$. In the middle row of the figure, the feasible sets F_r , F_s and F_t of the (extended) local systems of constraints corresponding with the cliques Cl_r , Cl_s and Cl_t , respectively, are shown. Recall that after Algorithm 4.48 has been applied an equilibrium as stated in Lemma 4.49 has been reached. In the figure, the symbol $=^i$ is used to denote that this equilibrium property holds between the indicated sets. In the sequel, we take this equilibrium property to be the *invariant* under evidence propagation, that is, if an equilibrium holds between the feasible sets of two local systems of constraints then an equilibrium has to hold between the feasible sets of the ‘updated’ systems of constraints.

Now suppose that we obtain the case-dependent evidence that the statistical

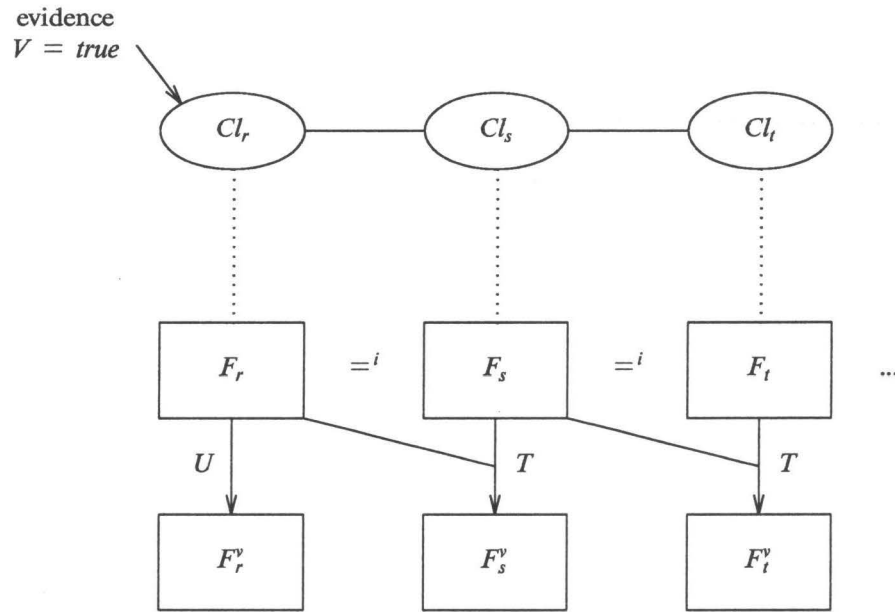


FIGURE 4.10. Propagating evidence.

variable V occurring in clique Cl_r has the value *true* (the case where we observe that V has the value *false* is dealt with analogously). For processing this evidence in Cl_r , we simply apply the method described in the preceding subsection: given the feasible set F_r of the (extended) local system of constraints for clique Cl_r , we compute the set $F_r^v = U(F_r)$, where U is the update mapping defined in Definition 4.51 (for ease of exposition, we assume in the remainder of this section that for each $y \in F_r$, the image $U(y) \neq \text{undefined}$). This updating is shown in the leftmost column of Figure 4.10. Informally speaking, each marginal distribution in F_r^v is the result of updating a marginal distribution from F_r . Recall that this set F_r^v of updated marginal distributions is a convex set and therefore can be described again by a system of linear constraints. It will be evident that in general the equilibrium property which held between the sets F_r and F_s , will no longer hold between the sets F_r^v and F_s .

For propagating the observed evidence from clique Cl_r to clique Cl_s , we now have to find a mapping T such that

- (1) given the feasible set F_s of the (extended) local system of constraints for clique Cl_s , the image F_s^v of F_s under T is the set of marginal distributions μ_{Cl_s} being extensions of the partial specification m_{Cl_s} associated with Cl_s after updating, and
- (2) the equilibrium property again holds between the sets F_r^v and F_s^v .

Furthermore, to render the notion of a partially quantified belief network invariant under evidence propagation, the image F_s^v of F_s under T has to be a convex set: only then is it possible to describe F_s^v by a system of linear constraints. So in addition T has to be a convex mapping.

In order to find such a mapping T , we reconsider the scheme for evidence propagation for the fully specified case presented by Lauritzen and Spiegelhalter. Let M be a specific global extension of the set M of partial specifications of marginal distributions of the network B . For the marginal distributions $\mu_{C_{l_i}}, \mu_{C_{l_s}} \in M$ we have that for each configuration $c_{C_{l_i} \cap C_{l_s}}$ of $C_{l_i} \cap C_{l_s}$ the property $\mu_{C_{l_i}}(c_{C_{l_i} \cap C_{l_s}}) = \mu_{C_{l_s}}(c_{C_{l_i} \cap C_{l_s}})$ holds. Now, let $\mu_{C_{l_i}}^v$ be the result of updating $\mu_{C_{l_i}}$ with the observed evidence. Recall from Definition 3.53 that the updated marginal distribution $\mu_{C_{l_i}}^v$ is computed using

$$\mu_{C_{l_i}}^v(C_{l_i}) = \mu_{C_{l_i}}(C_{l_i}) \cdot \frac{\mu_{C_{l_i}}^v(C_{l_i} \cap C_{l_s})}{\mu_{C_{l_i}}(C_{l_i} \cap C_{l_s})}$$

We reformulate this updating in terms of vectors from F_r and F_s . Let x be the vector from F_s defining the marginal distribution $\mu_{C_{l_i}} \in M$ we have considered before. Furthermore, let y be the vector from F_r defining the marginal distribution $\mu_{C_{l_s}} \in M$ corresponding with $\mu_{C_{l_i}}$. From the Lemmas 4.18 and 4.19 we have that for each configuration $c_{C_{l_i} \cap C_{l_s}}$ of $C_{l_i} \cap C_{l_s}$ there exist an index set \mathcal{J}_c such that

$$\mu_{C_{l_i}}(c_{C_{l_i} \cap C_{l_s}}) = \sum_{i \in \mathcal{J}_c} y_i$$

and an index set \mathcal{J}_e such that

$$\mu_{C_{l_s}}(c_{C_{l_i} \cap C_{l_s}}) = \sum_{i \in \mathcal{J}_e} x_i$$

Note that the index sets \mathcal{J}_e and \mathcal{J}_c are dependent upon the configuration $c_{C_{l_i} \cap C_{l_s}}$ under consideration. Now, recall that for each configuration $c_{C_{l_i} \cap C_{l_s}}$ of $C_{l_i} \cap C_{l_s}$ we have $\mu_{C_{l_i}}(c_{C_{l_i} \cap C_{l_s}}) = \mu_{C_{l_s}}(c_{C_{l_i} \cap C_{l_s}})$; so, we have

$$\sum_{i \in \mathcal{J}_e} x_i = \sum_{i \in \mathcal{J}_c} y_i$$

for each pair of appropriate index sets \mathcal{J}_e and \mathcal{J}_c . It will now be evident that we have to map each component $x_j = \mu_{C_{l_i}}(c_{C_{l_i}})$, $j = 1, \dots, 2^s$, of x into

$$\mu_{C_{l_i}}(c_{C_{l_i}}) \cdot \frac{\mu_{C_{l_i}}^v(c_{C_{l_i} \cap C_{l_s}})}{\mu_{C_{l_i}}(c_{C_{l_i} \cap C_{l_s}})} = x_j \cdot \frac{\sum_{i \in \mathcal{J}_j} U(y)_i}{\sum_{i \in \mathcal{J}_j} y_i}$$

where $U(y)_i$ denotes the i -th component of the image $U(y)$ of y and \mathcal{J}_j is the index set dependent upon x_j .

In the following definition we define a binary mapping T for this updating.

DEFINITION 4.65. *Let the update mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.51. We define the mapping $T: \mathbb{R}^{2^r} \times \mathbb{R}^{2^s} \rightarrow \mathbb{R}^{2^r}$, $s \geq 1$, by*

- $$(1) \quad T(x,y)_j = x_j \cdot \frac{\sum_{i \in \mathcal{J}_j} U(y)_i}{\sum_{i \in \mathcal{J}_j} y_i}, \text{ for } j = 1, \dots, 2^s, \text{ if } x \text{ and } y \text{ correspond as}$$
- indicated in the preceding discussion, and
- $$(2) \quad T(x,y) = \text{undefined, otherwise.}$$

Informally speaking, for a vector $x \in F_s$ and a corresponding vector $y \in F_r$, we have that some of the components of x are multiplied by a certain normalization factor dependent upon the chosen y , that another part of the components is multiplied by another normalization factor and so on.

The following lemma now states the evident property that when applied to vectors representing marginal distributions the mapping T yields a vector which again represents a marginal distribution, provided of course that the image is defined.

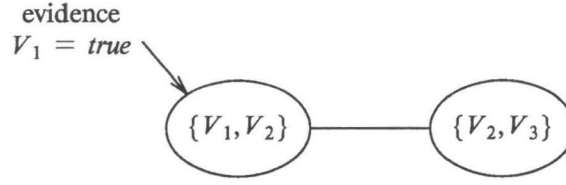
LEMMA 4.66. *Let the mapping $T: \mathbb{R}^{2^r} \times \mathbb{R}^{2^s} \rightarrow \mathbb{R}^{2^r}$, $r, s \geq 1$, be defined as above. Then, for each $x \in S_{2^r}$ and $y \in S_{2^s}$ we have that either $T(x,y) \in S_{2^r}$ or $T(x,y) = \text{undefined}$.*

The following lemma will be evident.

LEMMA 4.67. *Let the update mapping $U: \mathbb{R}^{2^r} \rightarrow \mathbb{R}^{2^r}$, $r \geq 1$, be defined according to Definition 4.51. Furthermore, let the mapping $T: \mathbb{R}^{2^r} \times \mathbb{R}^{2^s} \rightarrow \mathbb{R}^{2^r}$, $s \geq 1$, be defined as in Definition 4.65. Then, for each $x \in S_{2^r}$ and $y \in S_{2^s}$ such that $T(x,y) \neq \text{undefined}$, we have*

$$\sum_{i \in \mathcal{J}_j} T(x,y)_i = \sum_{i \in \mathcal{J}_j} U(y)_i \text{ for all } j = 1, \dots, 2^s.$$

The property from the preceding lemma guarantees that the equilibrium property holds between the two sets F_r^y and $F_s^x = \{T(x,y) \mid x \in F_s, y \in F_r, T(x,y) \neq \text{undefined}\}$. The mapping T therefore satisfies the first two properties we required T to have. Recall, however, that in addition T had to be a convex mapping. Unfortunately, the mapping T is not a convex one as the following easy counterexample will demonstrate.

FIGURE 4.11. The clique tree of G .

EXAMPLE 4.68. Let $B = (B, M)$ be a partially quantified belief network where G is a decomposable graph having the clique tree shown in Figure 4.11. For clique Cl_1 having the vertex set $V(Cl_1) = \{V_1, V_2\}$ we have the constituent probabilities

$$\begin{aligned} y_1 &= \mu_{Cl_1}(v_1 \wedge v_2) \\ y_2 &= \mu_{Cl_1}(v_1 \wedge \neg v_2) \\ y_3 &= \mu_{Cl_1}(\neg v_1 \wedge v_2) \\ y_4 &= \mu_{Cl_1}(\neg v_1 \wedge \neg v_2) \end{aligned}$$

where μ_{Cl_1} is a yet unknown marginal distribution associated with Cl_1 . For clique Cl_2 having the vertex set $V(Cl_2) = \{V_2, V_3\}$ we have the constituent probabilities

$$\begin{aligned} x_1 &= \mu_{Cl_2}(v_2 \wedge v_3) \\ x_2 &= \mu_{Cl_2}(\neg v_2 \wedge v_3) \\ x_3 &= \mu_{Cl_2}(v_2 \wedge \neg v_3) \\ x_4 &= \mu_{Cl_2}(\neg v_2 \wedge \neg v_3) \end{aligned}$$

where μ_{Cl_2} is a yet unknown marginal distribution associated with Cl_2 . Now suppose that after application of Algorithm 4.48, the feasible set F_{Cl_1} of the (extended) local system of constraints for Cl_1 is the line segment between the points z_1 and z_2 where

$$\begin{aligned} z_1 &= \left(\frac{1}{4}, \frac{1}{4}, 0, \frac{1}{2}\right), \text{ and} \\ z_2 &= \left(\frac{3}{4}, 0, 0, \frac{1}{4}\right) \end{aligned}$$

It will be evident that

$$z_3 = \left(\frac{1}{2}, \frac{1}{8}, 0, \frac{3}{8}\right)$$

lies on this line segment. Furthermore, suppose that the feasible set F_{Cl_2} of the (extended) local system of constraints for Cl_2 is the line segment between the points w_1 and w_2 where

$$w_1 = (0, \frac{1}{2}, \frac{1}{4}, \frac{1}{4})$$

$$w_2 = (\frac{1}{2}, 0, \frac{1}{4}, \frac{1}{4})$$

The point

$$w_3 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$$

lies on this line segment.

Suppose that we obtain the case-dependent evidence that $V_1 = \text{true}$. The images of z_1 , z_2 and z_3 under the update mapping U are computed as follows:

$$U(z_1) = (\frac{1}{2}, \frac{1}{2}, 0, 0)$$

$$U(z_2) = (1, 0, 0, 0)$$

$$U(z_3) = (\frac{4}{5}, \frac{1}{5}, 0, 0)$$

The reader can easily verify that the point $U(z_3)$ lies on the line segment between $U(z_1)$ and $U(z_2)$. We now consider the updating of w_1 , w_2 and w_3 . We choose z_1 to correspond with w_1 , z_2 to correspond with w_2 and z_3 to correspond with w_3 , and compute the images under T as follows:

$$T(w_1, z_1) = (0, \frac{1}{3}, \frac{1}{2}, \frac{1}{6})$$

$$T(w_2, z_2) = (\frac{2}{3}, 0, \frac{1}{3}, 0)$$

$$T(w_3, z_3) = (\frac{2}{5}, \frac{1}{10}, \frac{2}{5}, \frac{1}{10})$$

Upon inspection of these images it will be evident that $T(w_3, z_3)$ does not lie on the line segment between $T(w_1, z_1)$ and $T(w_2, z_2)$. It follows that the update mapping T is not convex. ■

From the previous example we have that the convex set of marginal distributions being extensions of a partially specified marginal distribution is mapped into a non-convex set by T : the local updating scheme presented by Lauritzen and Spiegelhalter evidently is a non-convex mapping.

To conclude, we recall that a partially quantified belief network can only be employed as a model for reasoning with uncertainty if it has associated two methods: a method for deriving information concerning probabilities of interest from the network and a method for propagating evidence through the network. From the foregoing discussion it will be evident that we have succeeded only in devising a method for the first goal; note that our method

for computing probability intervals from a partially quantified belief network allows for stepwise filling in the quantitative part of a belief network and therefore can be used as a help in the process of knowledge acquisition. We have not been able to find a method for evidence propagation that renders the notion of a partially quantified belief network and the linear programming approach invariant; in fact, our counterexample shows that for devising such a method we cannot build on the work by Lauritzen and Spiegelhalter.

References

- [ADAM84] J.B. ADAMS (1984). Probabilistic reasoning and certainty factors, in: B.G. BUCHANAN, E.H. SHORTLIFFE (eds). *Rule-Based Expert Systems. The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, Massachusetts, pp. 263 - 271.
- [ANDE89] S.K. ANDERSEN, K.G. OLESEN, F.V. JENSEN, F. JENSEN (1989). HUGIN — A shell for building Bayesian belief universes for expert systems, to appear in: *Proceedings of the 11th International Joint Conference of Artificial Intelligence*.
- [BERG73] C. BERGE (1973). *Graphs and Hypergraphs*, North-Holland, Amsterdam.
- [BIRK77] G. BIRKHOFF, S. MACLANE (1977). *A Survey of Modern Algebra*, MacMillan Publishing Company, NewYork.
- [BOOL54] G. BOOLE (1854). *An Investigation of the Laws of Thought, on which are founded the Mathematical Theories of Logic and Probabilities*, Walton and Maberley, London (reprinted in 1951 by Dover Publications, Inc., New York).
- [BRØN83] A. BRØNDSTED (1983). *An Introduction to Convex Polytopes*, Springer-Verlag (Graduate Texts in Mathematics; 90), New-York.
- [BUCH83] B.G. BUCHANAN, R.O. DUDA (1983). Principles of rule-based expert systems, *Advances in Computers*, vol. 22, pp. 163 - 216.

- [BUCH84] B.G. BUCHANAN, E.H. SHORTLIFFE (1984). *Rule-Based Expert Systems. The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, Massachusetts.
- [CARN50] R. CARNAP (1950). *Logical Foundations of Probability*, University of Chicago Press, Chicago.
- [CHEE85] P. CHEESEMAN (1985). In defense of probability, *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pp. 1002 - 1009.
- [CHEE88] P. CHEESEMAN (1988). An inquiry into computer understanding, *Computational Intelligence*, vol. 4, pp. 58 - 66.
- [COHE85] P.R. COHEN (1985). *Heuristic Reasoning About Uncertainty: An Artificial Intelligence Approach*, Pitman, London.
- [COOP87] G.F. COOPER (1987). *Probabilistic Inference Using Belief Networks is NP-Hard*, Memo KLS-87-27, Knowledge Systems Laboratory, Medical Computer Science Group, Stanford University, Stanford, California.
- [COX79] R.T. COX (1979). Of inference and inquiry — an essay in inductive logic, in: L. LEVINE, M. TRIBUS (eds). *The Maximum Entropy Formalism*, MIT Press, Cambridge, Massachusetts.
- [DARR80] J.N. DARROCH, S.L. LAURITZEN, T.P. SPEED (1980). Markov fields and log-linear interaction models for contingency tables, *The Annals of Statistics*, vol. 8, pp. 522 - 539.
- [DECH87] R. DECHTER, J. PEARL (1987). *Tree Clustering Schemes for Constraint Processing*, Report CSD-870054, University of California, Los Angeles.
- [DEMP68] A.P. DEMPSTER (1968). A generalisation of Bayesian inference, *Journal of the Royal Statistical Society (Series B)*, vol. 30, pp. 205 - 247.
- [DEMP88] A.P. DEMPSTER, A. KONG (1988). Uncertain evidence and artificial analysis, *Journal of Statistical Planning and Inference*, vol. 20, pp. 355 - 368.
- [DOMB72] F.T. DE DOMBAL, D.J. LEAPER, J.R. STANILAND, A.P. MCCANN, J.C. HORROCKS (1972). Computer-aided diagnosis of acute abdominal pain, *British Medical Journal*, vol. 2, pp. 9 - 13.
- [DOMB74] F.T. DE DOMBAL, D.J. LEAPER, J.C. HORROCKS, J.R. STANILAND, A.P. MCCANN (1974). Human and computer-aided diagnosis of abdominal pain: further report with emphasis on the performance of clinicians, *British Medical Journal*, vol. 4, pp. 376 - 380.

- [DUDA76] R.O. DUDA, P.E. HART, N.J. NILSSON (1976). *Subjective Bayesian Methods for Rule-Based Inference Systems*, Technical Note 124, Artificial Intelligence Center, SRI International, Menlo Park.
- [DUDA79] R.O. DUDA, J.G. GASCHNIG, P.E. HART (1979). Model design in the PROSPECTOR consultant system for mineral exploration, in: D. MICHIE (ed). *Expert Systems in the Micro Electronic Age*, Edinburgh University Press, Edinburgh, pp. 153 - 167.
- [FINE70] B. DE FINETTI (1970). *Theory of Probability*, Wiley, New York.
- [GORR68] G.A. GORRY, G.O. BARNETT (1968). Experience with a model of sequential diagnosis, *Computers and Biomedical Research*, vol. 1, pp. 490 - 507.
- [HAIL65] T. HAILPERIN (1965). Best possible inequalities for the probability of a logical function of events, *American Mathematical Monthly*, vol. 72, pp. 343 - 359.
- [HAIL86] T. HAILPERIN (1986). *Boole's Logic and Probability*, 2nd Edition, North-Holland, Amsterdam.
- [HECK86] D.E. HECKERMAN (1986). Probabilistic interpretations for MYCIN's certainty factors, in: L.N. KANAL, J.F. LEMMER (eds). *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, pp. 167 - 196.
- [HENR89] M. HENRION (1989). Some practical issues in constructing belief networks, in: L.N. KANAL, T.S. LEVITT, J.F. LEMMER (eds). *Uncertainty in Artificial Intelligence 3*, North-Holland, Amsterdam, pp. 161 - 173.
- [HOGG78] R.V. HOGG, A.T. CRAIG (1978). *Introduction to Mathematical Statistics*, MacMillan Publishing Company, New York.
- [HORV86] E.J. HORVITZ, D.E. HECKERMAN, C.P. LANGLOTZ (1986). A framework for comparing alternative formalisms for plausible reasoning, *Proceedings of the 5th National Conference on Artificial Intelligence*, pp. 210 - 214.
- [HORV88] E.J. HORVITZ, J.S. BREESE, M. HENRION (1988). Decision theory in expert systems and artificial intelligence, *International Journal of Approximate Reasoning*, vol. 2, pp. 247 - 302.
- [ISHI81] M. ISHIZUKA, K.-S. FU, J.T.P. YAO (1981). *A Theoretical Treatment of Certainty Factor in Production Systems*, Report no. CE-STR-81-6, School of Civil Engineering, Purdue University, West Lafayette.
- [JENS88a] F.V. JENSEN (1988). *Junction Trees and Decomposable Hypergraphs*, JUDEX Research Report, Aalborg.

- [JENS88b] F.V. JENSEN, K.G. OLESEN, S.K. ANDERSEN (1988). *An Algebra of Bayesian Belief Universes for Knowledge Based Systems*, Report R-88-25, Institute of Electronic Systems, Aalborg University, Aalborg.
- [KAHN82] D. KAHNEMAN, P. SLOVIC, A. TVERSKY (1982). *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
- [KANA86] L.N. KANAL, J.F. LEMMER (1986). *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam.
- [KANA89] L.N. KANAL, T.S. LEVITT, J.F. LEMMER (1989). *Uncertainty in Artificial Intelligence 3*, North-Holland, Amsterdam.
- [KIIV84] H. KIIVERI, T.P. SPEED, J.B. CARLIN (1984). Recursive causal models, *Journal of the Australian Mathematical Society (Series A)*, vol. 36, pp. 30 - 52.
- [KIM83] J.H. KIM, J. PEARL (1983). A computational model for causal and diagnostic reasoning in inference systems, *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pp. 190 - 193.
- [LASD70] L.S. LASDON (1970). *Optimization Theory for Large Systems*, MacMillan Publishing Company, New York.
- [LAUR84] S.L. LAURITZEN, T.P. SPEED, K. VIJAYAN (1984). Decomposable graphs and hypergraphs, *Journal of the Australian Mathematical Society (Series A)*, vol. 36, pp. 12 - 19.
- [LAUR88a] S.L. LAURITZEN, D.J. SPIEGELHALTER (1987). Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal of the Royal Statistical Society (Series B)*, vol. 50, no. 2, pp. 157 - 224.
- [LAUR88b] S.L. LAURITZEN, A.P. DAWID, B.N. LARSEN, H.-G. LEIMER (1988). *Independence Properties of Directed Markov Fields*, Report R-88-32, Institute of Electronic Systems, Aalborg University, Aalborg, Denmark.
- [LEMM88] J.F. LEMMER, L.N. KANAL (1988). *Uncertainty in Artificial Intelligence 2*, North-Holland, Amsterdam.
- [LIND79] D.V. LINDLEY, A. TVERSKY, R.V. BROWN (1979). On the reconciliation of probability assessments, *Journal of the Royal Statistical Society, Series A*, vol. 142, part 2, pp. 146 - 180.
- [LUCA90] P.J.F. LUCAS, L.C. VAN DER GAAG (1990). *Principles of Expert Systems*, Addison Wesley, to appear.
- [MAIE83] D. MAIER (1983). *The Theory of Relational Databases*, Pitman, London.

- [MALC86] A. MALCHOW-MØLLER, C. THOMSEN, P. MATZEN, L. MINDEHOLM, B. BJERREGAARD, S. BRYANT, J. HILDEN, J. HOLST-CHRISTENSEN, T.S. JOHANSEN, E. JUHL (1986). Computer diagnosis in jaundice. Bayes' Rule founded on 1002 consecutive cases, *Journal of Hepatology*, vol. 3, pp. 154 - 163.
- [McDE80] D. McDERMOTT, J. DOYLE (1980). Non-monotonic logic I, *Artificial Intelligence*, vol. 13, pp. 41 - 72.
- [NEAP90] R.E. NEAPOLITAN (1990). *Probabilistic Reasoning in Expert Systems. Theory and Algorithms*, John Wiley & Sons, New York.
- [PAPA82] C.H. PAPADIMITRIOU, K. STEIGLITZ (1982). *Combinatorial Optimization. Algorithms and Complexity*, Prentice-Hill, Englewood Cliffs, New Jersey.
- [PEAR85] J. PEARL (1985). *How To Do With Probabilities What People Say You Can't*, Report CSD-850031, University of California, Los Angeles.
- [PEAR88] J. PEARL (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Palo Alto.
- [PITM76] J.W. PITMAN (1976). *Markov Random Fields*, Unpublished Lecture Notes from the University of Copenhagen.
- [REBA89] G. REBANE, J. PEARL (1989). The recovery of causal poly-trees from statistical data, in: L.N. KANAL, T.S. LEVITT, J.F. LEMMER (eds). *Uncertainty in Artificial Intelligence 3*, North-Holland, Amsterdam, pp. 175 - 182.
- [SAVA54] L.J. SAVAGE (1954). *The Foundations of Statistics*, Wiley, New York.
- [SCHR86] A. SCHRIJVER (1986). *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester.
- [SHAC86] R.D. SHACHTER (1986). Intelligent probabilistic inference, in: L.N. KANAL, J.F. LEMMER (eds). *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, pp. 371 - 382.
- [SHAF76] G. SHAFER (1976). *A Mathematical Theory of Evidence*, Princeton University Press, Princeton.
- [SHEN86] P. SHENOY, G. SHAFER (1986). Propagating belief functions with local computations, *IEEE Expert*, vol. 1, pp. 43 - 52.

- [SHOR84] E.H. SHORTLIFFE, B.G. BUCHANAN (1984). A model of inexact reasoning in medicine, in: B.G. BUCHANAN, E.H. SHORTLIFFE (eds). *Rule-Based Expert Systems. The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, Massachusetts, pp. 233 - 262.
- [SPIE86a] D.J. SPIEGELHALTER (1986). A statistical view of uncertainty in expert systems, in: W. GALE (ed). *Artificial Intelligence and Statistics*, Reading, Massachusetts, Addison-Wesley, pp. 17 - 55.
- [SPIE86b] D.J. SPIEGELHALTER (1986). Probabilistic reasoning in predicative expert systems, in: L.N. KANAL, J.F. LEMMER (eds). *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, pp. 47 - 67.
- [STRA76] G. STRANG (1976). *Linear Algebra and Its Applications*, Academic Press, New York.
- [SZOL78] P. SZOLOVITS, S.G. PAUKER (1978). Categorical and probabilistic reasoning in medical diagnosis, *Artificial Intelligence*, vol. 11, pp. 115 - 144.
- [TARJ84] R.E. TARJAN, M. YANNAKAKIS (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs, *SIAM Journal of Computing*, vol. 13, no. 3, pp. 566 - 579.
- [TRUN88] L.M. TRUNG (1988). The order of node removals in influence diagrams, *Proceedings of the Workshop on Uncertainty Processing in Expert Systems*, Alšovice, Csechoslovakia, pp. 20 - 24.
- [WARN61] H.R. WARNER, A.F. TORONTO, L.G. VEASY, R. STEPHENSON (1961). A mathematical approach to medical diagnosis: application to congenital heart disease, *Journal of the American Medical Association*, vol. 177, pp. 177 - 183.
- [WERM83] N. WERMUTH, S.L. LAURITZEN (1983). Graphical and recursive models for contingency tables, *Biometrika*, vol. 70, pp. 537 - 552.
- [WILS79] R.J. WILSON (1979). *Introduction to Graph Theory*, 2nd Edition, Longman, Harlow, Essex.
- [WISE86] B.P. WISE, M. HENRION (1986). A framework for comparing uncertain inference systems to probability, in: L.N. KANAL & J.F. LEMMER (eds). *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, pp. 69 - 83.
- [YANN81] M. YANNAKAKIS (1981). Computing the minimum fill-in is NP-complete, *SIAM Journal of Algebraic Discrete Methods*, vol. 2, pp. 77 - 79.

- [ZADE78] L.A. ZADEH (1978). Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets & Systems*, vol. 1, pp. 3 - 28.
- [ZADE83] L.A. ZADEH (1983). The role of fuzzy logic in the management of uncertainty in expert systems, *Fuzzy Sets & Systems*, vol. 11, pp. 199 - 228.
- [ZADE86] L.A. ZADEH (1986). Is probability theory sufficient for dealing with uncertainty in AI: a negative view, in: L.N. KANAL, J. LEMMER (eds). *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, pp. 103 - 116.

Samenvatting

Bij het ontwikkelen van een kennissysteem voor een bepaald probleemgebied blijkt vaak dat (menselijke) experts op het betreffende gebied oordelen kunnen vormen en beslissingen kunnen nemen op grond van onvolledige, onzekere en soms zelfs tegenstrijdige informatie. Om toegepast te kunnen worden in zo'n probleemgebied moet een kennissysteem ook in staat zijn met dit soort informatie om te gaan. Onderzoek naar formalismen voor de representatie van onzekerheid en algoritmen voor het manipuleren van onzekere informatie vormt dan ook een substantieel onderzoeksgebied binnen de kunstmatige intelligentie. Dit proefschrift heeft de toepasbaarheid van de waarschijnlijkheidsrekening in een dergelijke context tot onderwerp.

Aangezien de waarschijnlijkheidsrekening één van de oudste theorieën met betrekking tot onzekerheid is, is het niet verwonderlijk dat deze wiskundige theorie in de jaren zeventig als eerste uitgangspunt in het onderzoek naar het redeneren met onzekerheid werd gekozen. Helaas blijkt een aantal problemen een naïeve toepassing van de waarschijnlijkheidsrekening in een kennissysteem in de weg te staan. Voor een traditionele besliskundige aanpak zijn bijvoorbeeld al exponentieel veel kansen nodig (exponentieel in het aantal onderscheiden statistische variabelen), meer dan gewoonlijk in praktische toepassingen bekend zullen zijn. Daarnaast is de tijdscomplexiteit van de algoritmen voor het uitvoeren van de probabilistische berekeningen voor het manipuleren van onzekere informatie exponentieel.

In de zeventiger jaren werd daarom gezocht naar modificaties van de waarschijnlijkheidsrekening die oplossingen moesten bieden voor de gesignaleerde problemen. De in deze periode ontwikkelde modellen gaan in beginsel uit van de waarschijnlijkheidsrekening, maar bieden ad hoc methoden voor bijvoorbeeld het manipuleren van een partiële specificatie van een kansverdeling. In dit proefschrift worden deze modellen quasi-probabilistische

modellen genoemd. De quasi-probabilistische modellen zijn vanuit wiskundig oogpunt helaas niet correct; desondanks worden ze in de praktijk op grote schaal toegepast. Met name deze observatie heeft een zorgvuldige analyse van één van deze modellen, te weten het certainty factor model, gemotiveerd. De resultaten hiervan zijn vastgelegd in hoofdstuk 2; aangetoond wordt bijvoorbeeld dat in het model impliciet sterke veronderstellingen betreffende onafhankelijkheid zijn gedaan.

De eerste teleurstellende ervaringen met het toepassen van de waarschijnlijkheidsrekening in kennissystemen hebben ertoe geleid dat de aandacht ervoor als theoretische basis voor het redeneren met onzekerheid drastisch afnam: de geschiktheid van de waarschijnlijkheidsrekening werd zelfs ter discussie gesteld. De tegenstanders van de waarschijnlijkheidsrekening wijzen erop dat deze theorie niet voor elke vorm van onzekerheid een natuurlijke oplossing biedt, en stellen alternatieve grondslagen voor het redeneren met onzekerheid voor, zoals de vage logica, de theorie van Dempster en Shafer, en een aantal niet-numerieke theorieën.

Het onderzoek naar de toepassing van de waarschijnlijkheidsrekening in kennissystemen is echter voortgezet en heeft de laatste jaren een nieuwe impuls gekregen in de vorm van de zogenaamde netwerkmodellen. Deze modellen worden gekenmerkt door een grafisch model van de statistische variabelen die in een probleemdomain onderscheiden worden en hun onderlinge probabilistische relaties; de sterkten van deze relaties zijn vastgelegd met behulp van voorwaardelijke kansen, die tesamen een unieke kansverdeling op het domein definiëren. Hoofdstuk 3 gaat op deze nieuwe stroming in het redeneren met onzekerheid in en bespreekt een voorbeeld van een dergelijk netwerkmodel in detail.

In tegenstelling tot de quasi-probabilistische modellen zijn de netwerkmodellen wiskundig correct. De netwerkmodellen zijn echter niet in staat om met een partiële specificatie van een kansverdeling om te gaan. In hoofdstuk 4 van dit proefschrift wordt als gedeeltelijke oplossing voor dit probleem een methode voorgesteld om, gegeven een partieel gekwantificeerd netwerk, grenzen aan kansen te berekenen; deze methode kan bijvoorbeeld als hulpmiddel bij het kwantificeren van de probabilistische relaties in een netwerk gebruikt worden. De voorgestelde methode lost het genoemde probleem echter slechts ten dele op: het is vooralsnog niet mogelijk om met een partieel gekwantificeerd netwerk te manipuleren.