CHRISTIAN BORGELT AND RUDOLF KRUSE

ABDUCTIVE INFERENCE WITH PROBABILISTIC NETWORKS

1 INTRODUCTION

Abduction is a form of non-deductive logical inference. Examples given by [Peirce, 1958], who is said to have coined the term "abduction", include the following:

> I once landed at a seaport in a Turkish province; and as I was walking up to the house which I was to visit, I met a man upon horseback, surrounded by four horsemen holding a canopy over his head. As the governour of the province was the only personage I could think of who would be so greatly honoured, I inferred that this was he. This was a hypothesis.
>
> Fossils are found; say remains like those of fishes, but far in the interior of the country. To explain the phenomenon, we suppose the sea once washed over this land. This is another hypothesis.
>
> Numberless documents and monuments refer to a conqueror called Napoleon Bonaparte. Though we have not seen him, what we have seen, namely all those documents and monuments, cannot be explained without supposing that he really existed. Hypothesis again.

On the other hand, probabilistic networks [Pearl, 1992] are a method to structure a multivariate probability distribution and to compute efficiently (conditioned) marginal distributions on subspaces. Hence, at first glance, there seems to be little connection between abduction and probabilistic networks. Therefore we start this chapter by showing how the two notions are connected through hypothesis assessment and statistical explanations. In addition, we use these starting sections to argue for a certain way of defining abductive inferences in contrast to inductive ones (although, unfortunately, this involves repeating some parts of what was already discussed in the introductory chapter). We believe that this view could help to avoid a lot of confusion that seems to prevail w.r.t. the term "abduction".

Next we discuss a general model of abductive inference. However, this model is not suited for implementation, because it needs too much storage space. Direct approaches to simplify the model render it manageable, but require strong independence assumptions that are hardly acceptable in applications. Therefore a modeling technique is desired, by which we can take into account dependences between the involved variables, but which nevertheless lets us exploit (conditional) independences to simplify the model. One such technique, which has become very popular nowadays, are probabilistic networks. We review this modeling technique and discuss how probabilistic networks can be used for abductive inference.

2 OUR CATEGORIZATION OF LOGICAL INFERENCES

Logic, in the most general sense, describes the structure of languages in which one can argue. That is, logic is the (formal) theory of arguments, where an *argument* is a group of statements that are related to each other. An argument consists of one statement representing the *conclusion* and one or more statements that give reasons supporting it. The latter statements are called *premisses* [Salmon, 1973]. The process of deriving the conclusion from the premisses (using an argument) is called an *inference*. Arguments are studied by analyzing the *inference rule* used to draw the inference. Such rules are usually stated in the form of *argument schemes*. For example, the well-known inference rule of *modus ponens* is characterized by the following argument scheme:

$$\begin{array}{l} A \rightarrow B \\ A \\ \hline B \end{array}$$

If a given argument can be constructed from this scheme by replacing A and B with suitable statements, then the *modus ponens* was used to draw the inference.

As already said in the introduction, abduction is a form of logical inference. Therefore, in this section we briefly study the specific properties of *abductive* inferences in contrast to *deductive* and *inductive* inferences (as we see them) by discussing their characteristic features. We do so, because our interpretation of the term “abduction” is a specific one that differs (although only slightly) from explications given by some other authors. Thus, the main aim of this section is to avoid confusion by introducing precise notions. For a more detailed discussion of other categorizations of logical inferences and especially of other criteria to distinguish between induction and abduction than those we use, see the introductory chapter.

2.1 Deduction and Reduction

Łukasiewicz showed (according to [Bochenski, 1954]), that all logical inferences can be divided into two classes, which he called *deduction* and *reduction*. The idea underlying this division is the following: By exploiting logical equivalences we can modify the premisses of all arguments in such a way that arguments with only two premisses result. One of these premisses is a *conditional* or an *implication* (an if-then-statement), the other is equivalent either to the antecedent or to the consequent of this conditional. Written as argument schemes these two cases look like this:

$$\text{Deduction:} \quad \begin{array}{l} A \rightarrow B \\ A \\ \hline B \end{array} \qquad\qquad \text{Reduction:} \quad \begin{array}{l} A \rightarrow B \\ B \\ \hline A \end{array}$$

(To avoid some technical problems, we implicitly assume throughout this chapter that conditionals may or may not be (multiply) universally quantified. This saves the (always possible) derivation of e.g. $A(c) \to B(c)$ from $\forall x : A(x) \to B(x)$ for some appropriate constant c.)

Obviously, both of these inference rules are based on the logical tautology $((A \to B) \wedge A) \to B$. However, they use it in different ways. The first scheme is the *modus ponens* in its usual form. It corresponds exactly to the tautology, since the inference is drawn in the direction of the implication from A to B. The second scheme looks like some kind of reversal of the first, since the inference is drawn in the opposite direction. In the following we briefly study the properties of these two kinds of inferences.

Deduction serves the purpose to make explicit all truths that are determined by a set of statements. We can find these truths by constructing appropriate deductive arguments. Deduction is the basis of the hypothetico-deductive method used in science [Hempel, 1966]: A set of statements, the so-called *axioms*, is fixed. Then one tries to find all consequences of the axioms using deductive inferences (in which conclusions derived earlier can also be used as premisses). In the natural sciences these consequences are eventually compared to experimental findings to check the validity of the axioms (which, if the predicted experimental results do not show up, are refuted by applying the *modus tollens*). (A more detailed discussion can be found in the section on Peirce's inferential theory in the introductory chapter.) For this method to be acceptable, it is necessary that deductive inferences yield only true statements provided that the premisses are true. This can be guaranteed only, if no information is added to the information already present in the premisses. (If information was added, we could not guarantee the truth of the conclusion, simply because we would not know whether the additional information is correct.) Obviously deduction fulfills these requirements. Thus, the basic properties of deduction are that it is infallible, but it does not tell us anything new (w.r.t. the premisses). These properties are a consequence of the fact that deductive inferences corresponds exactly to tautologies.

Reduction serves the purpose to find explanations for statements that describe e.g. observations made. Obviously, the second premiss of a reductive argument can be obtained from the first premiss (the conditional) and the conclusion by a deductive inference. This is the rationale underlying reductive arguments: The premiss B becomes a logical consequence and is thus "explained" by the conclusion (provided the other premiss, i.e., the conditional, is given). However, there is a serious drawback: Reductive inferences are not truth preserving. Indeed, this type of inference is well-known in (deductive) logic as the fallacy of confirming the consequent [Salmon, 1973]. The conclusion may be false, even if both premisses are true. This is not surprising, since information about the statement A is added, namely, that it is not possible that A is false if B is true (although the conditional would be true in this case). Obviously, this additional information could be false. Thus, the basic properties of reduction are that it is fallible, but as a compensation it tells us something new (w.r.t. the premisses). These properties are a consequence of the fact that there is no tautology to which reductive inferences correspond.

2.2 Induction and Abduction

Often abduction is defined as what we (following Łukasiewicz) called reduction. All explanatory inferences, that is, all inferences yielding conclusions from which one of the premisses can be derived deductively (given the other premiss) are then called abductive. This approach is closely connected to the tradition of calling all non-deductive arguments inductive arguments [Salmon, 1973], which seems to go back to Aristotle [Losee, 1993] (see section 2.1 of the introductory chapter for a more detailed discussion of this view).

Another approach, which traces back to [Peirce, 1958], contrasts induction and abduction by associating them with different argument schemes, namely:

$$\text{Induction:}\quad \begin{array}{l} A \\ B \\ \hline A \rightarrow B \end{array} \qquad\qquad \text{Abduction:}\quad \begin{array}{l} A \rightarrow B \\ B \\ \hline A \end{array}$$

(Obviously it can be argued that both schemes are based on the aforementioned tautology $((A \rightarrow B) \wedge A) \rightarrow B$.) More details on this view can be found in section 2.2 of the introductory chapter of this book.

Although we are more sympathetic to the second approach, we consider both to be unacceptable. The first we reject, because it does not provide grounds on which to distinguish between induction and abduction (see also below). However, the two notions are not used interchangeably. In practice a distinction is made (although it is often rather vague).

The second approach we reject because of the strange form given to the argument scheme of inductive inference, since it leads to all kinds of problems. In the first place, where does the conclusion come from? It is not part of the premisses as in the other two argument schemes. Why is "$A \rightarrow B$" the conclusion and not "$B \rightarrow A$"? (Obviously this depends on the order of the premisses, i.e., on something that is irrelevant to the other two schemes.) If the resulting conditional is (universally) quantified[1], but the premisses are not, how are the constants chosen that are to be replaced by variables? However, a more important objection is the following: Łukasiewicz showed (according to [Bochenski, 1954]) that induction is only a special case of reduction. A simple example (taken from [Bochenski, 1954]) will make this clear: Let us assume that we experimented with three pieces of phosphorus, a, b, and c, and found that they caught fire below 60^oC. We infer that all pieces of phosphorus behave in this fashion. What is the argument scheme of this inference? Obviously it is the following:

$$\begin{array}{l} \text{If all pieces of phosphorus catch fire below } 60^oC, \\ \quad \text{then } a, b, \text{ and } c \text{ will.} \\ a, b, \text{ and } c \text{ catch fire below } 60^oC. \\ \hline \text{All pieces of phosphorus catch fire below } 60^oC. \end{array}$$

[1] As already said above, we implicitly assume that all conditionals may or may not be (multiply) universally quantified.

Clearly, this is a reduction. Thus, induction and abduction, the latter of which is characterized by the reductive argument scheme in the second approach, would be indistinguishable. Or, to put it differently: It sounds unreasonable to call the same inference once an induction and once an abduction depending on the form of the argument scheme, if the schemes can be transformed into each other. Note also, that seeing induction as a special case of reduction removes the problem how we arrive at the special form of the conclusion: It is a part of the first premiss. Of course, we now face the problem where the first premiss comes from. However, it is not a problem of logic where the premisses come from. In logic they are always assumed to be given and it is merely studied how the conclusion is derived from them. We see it as an advantage that the problem of how to generate the hypotheses that enter the inferences is removed from the realm of logic. (Describing induction as an inference leading from A and B to $A \rightarrow B$ is a trial to incorporate the hypothesis generation process into logic.)[2]

In contrast to the approaches discussed above, our own distinction between induction and abduction refers to [Popper, 1934][3]:

> An inference is usually called an inductive inference or an induction, if it goes from *particular statements*, which describe e.g. observations, experiments, etc., to *general statements*, i.e., to hypotheses and theories.

That is, we do not base our distinction on the argument scheme used, since for both induction and abduction it is the reductive scheme, but on the type of the statement inferred, which indeed differs for these two types of inferences.

Note that this distinction is not contained in the second approach discussed above. If we conclude from A and B that $A \rightarrow B$, we will not necessarily have inferred a general statement. If A and B are particular statements, then $A \rightarrow B$ is also a particular statement (provided no quantifier is added). That it is often seen as a general statement nevertheless is due to the fact that general laws often come in the form of conditional statements. However, that a statement is a *law*, i.e., that it is valid in general, is expressed by variables and quantifiers. Of course, we could require the conclusion to be universally quantified, but, as already indicated, which constants are to be replaced by variables is completely arbitrary and not determined by logical rules.

With induction being an inference leading to a general statement, it is natural to define abduction as its counterpart [Borgelt, 1992]:

> Abduction is a reductive inference in which the conclusion is a particular statement.

Using a notion of formal logic, we may also say that abduction is a reductive inference with the conclusion being a *ground formula*. That is, the conclusion must

[2] Thus we place ourselves in the tradition that distinguishes the *context of discovery* and the *context of justification* of (scientific) hypotheses [Losee, 1993] and declare that logic belongs to the latter.

[3] Our translation from the German, italics in the original.

not contain variables, neither bound nor free. Using a term introduced in the introductory chapter, we can say that we define abduction by requiring that the set of *abducibles* must contain only ground formulae of a given formal language. We do not admit existentially quantified formulae (as some other authors do), since $\exists x : A(x)$ is equivalent to $\neg\forall x : \neg A(x)$ and thus tantamount to a general statement. By the restriction to ground formulae we want to make explicit that specific facts are inferred and *not* general laws (for which we need variables and quantifiers). In contrast to this, induction infers general laws and *not* specific facts.[4] Hence, with induction one generalizes observations, with abduction one learns about unobserved or even unobservable (particular) facts.

One may ask why an analogous distinction is not made for deductive inferences. To this we reply that such a distinction *is* made in the theory of the hypothetico-deductive method of science. This distinction is necessary, because only particular statements can be confronted with experimental findings. We can never observe directly a general law [Popper, 1934]. Thus we need to distinguish inferences which, for example, derive statements of lower generality from statements of higher generality (in order to find more specific theories) from inferences which yield particular statements (which can be tested in experiments). However, as far as we know, there are no special names for these two types of deductive inferences and thus this distinction may slip attention. Unfortunately a closer inspection of this astonishingly far reaching analogy is beyond the scope of this paper.

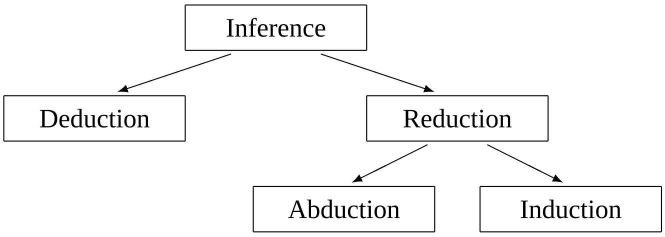


Figure 1. Categories of logical inferences.

The classification scheme for logical inferences we arrived at is shown in figure 1. The distinction between abductive and reductive inferences is based on the form of the argument scheme, the distinction between inductive and abductive inferences is based on the form of the conclusion. This scheme looks very much like the one discussed in the section on Peirce's syllogistic theory in the introductory chapter. However, it should be noted that the distinction in the non-deductive branch is based on a different criterion and hence the two schemes are not identical.

[4] It can be seen from the examples given by [Peirce, 1958] that he must have had something like this in mind, since in his examples of abductive arguments the result is a particular statement, see e.g. the examples cited in the introduction. However, he may have permitted existentially quantified formulae.

3 HYPOTHESIS ASSESSMENT

As already mentioned above, a reductive—and thus an abductive—inference can yield a false conclusion, even if the premisses are true. This is usually made explicit by calling the conclusion of a reductive argument a *hypothesis* or a *conjecture*. Of course, we do not want our conclusions to be wrong. Therefore we look for criteria to assess the hypotheses obtained by reductive arguments in order to minimize the chances that they are wrong. This is a perfectly reasonable intention. However, it has lead to fairly unpleasant situation: One of the main problems why there is such confusion about what the term "abduction" means is the fact that nearly nobody distinguishes clearly between the logical inference and the assessment of its result. This can already be seen from a frequently used characterization of abduction, namely, that it is an "inference to the best(!) explanation" (cf., for example, [Josephson and Josephson, 1996]). That is, not all explanatory inferences (even if their conclusions are only ground formulae, see above) qualify as abductive inferences. To qualify, they must yield a "better" explanation than all other explanatory inferences (or at least an equally good one). However, it is clear that depending on the domain of application different criteria may be used to specify what is meant by "best". Thus it is not surprising that there are as many different interpretations of the term "abduction" as there are criteria to assess hypotheses. In the preceding section we tried to give a purely logical definition of the term *abduction*, which enables us to study now in a systematic way different criteria to assess hypotheses without introducing any ambiguity w.r.t. the term "abduction".

We do not claim our list of criteria to be complete, but only give some examples in order to make clear the main difficulties involved. (Note that for the most part of this section we do not distinguish between abduction and induction, since the criteria studied apply to all reductive arguments.) More details on hypothesis assessment can be found in the introductory chapter.

Relation between antecedent and consequent. If we require only that a reductive argument explains its second premise by yielding a hypothesis from which it can be derived deductively (given the conditional), a lot of useless or even senseless "explanations" are possible. For example, consider the conditional "If sparrows can fly, then snow is white." (which, from a purely logical point of view, is true due to the "material" implication). However, we are not satisfied with the explanation "sparrows can fly", if we try to figure out why snow is white, because we cannot see any connection between the two facts. The situation gets even worse, if we replace the antecedent by a contradiction, since: *ex contradictio quodlibet* (from a contradiction follows whatever you like). Therefore, from a logical point of view, a contradiction is a universal explanation.

The problem is that we want an explanation to give a reason, or even better, a cause for the fact to be explained.[5] However, "reason" and "cause" are semantical

[5] Here the difference between the if-then in logic and the if-then in natural language is revealed most clearly. In natural language most often a causal connection or a inherence relation is assumed between

notions, referring to the *meaning* of statements. Yet the meaning of statements cannot be captured by formal logic, which handles only the “truth functionality” of compound statements.

Relation to other statements. Usually drawing an inference is not an isolated process, but takes place in an “environment” of other statements, which are known or at least believed to be true. These other statements can (but need not) have a bearing on the conclusion of a reductive argument. They may refute or support it. For example, the fact that clouds are white may be explained by the hypothesis that they consist of cotton wool. However, we also know that cotton wool is heavier than air. Thus we can refute the hypothesis (although indirectly, since we need a (deductive) argument to make the contradiction explicit). On the other hand, if we observe white animals, we may conjecture that they are swans. This conjecture is supported, if we learn also that the animals can fly and have orange bills. That is, a hypothesis gets more plausible, if several reductive arguments lead to it.

Parsimony (Ockham’s razor). Not only do we want explanations to give reasons or causes and to be compatible with our background knowledge, we also want them to be as simple as possible. For example, suppose that your car does not start and that the headlights do not come on. Considering as explanations that both headlights are burned out *and* that the starter is broken or that simply the battery is empty, you would surely prefer the latter. The rationale of this preference is expressed in Ockham’s razor[6]: *pluralitas non est ponenda sine necessitate*, that is, multiple entities should not be assumed without necessity. (Note that this is also a semantic criterion, although it may be turned into a syntactical one by fixing the formal language and introducing a measure of complexity for the formulae.)

Probability. Consider the following two abductive arguments:

Water is (1) a liquid, (2) transparent, and (3) colorless.
This substance is (1) a liquid, (2) transparent, and (3) colorless.

This substance is water.

Tetrachloromethane is (1) a liquid, (2) transparent, and (3) colorless.
This substance is (1) a liquid, (2) transparent, and (3) colorless.

This substance is tetrachloromethane.

W.r.t. the criteria mentioned above they are equivalent. Nevertheless we prefer the first, since in daily life we deal with water much more often than with tetrachloromethane. Due to this difference in frequency we judge the first conclusion to be much more *probable* than the second. This is one way in which the notion of *probability* enters our discussion, namely as a means to assess hypotheses.

the antecedent and the consequent of a conditional. In formal logic, no such assumption is made.

[6] William of Ockham, 1280–1349.

4 PROBABILISTIC INFERENCES

Up to now we assumed implicitly that the conditional that appears in deductive as well as in reductive arguments is known to be absolutely correct. However, in real world applications we rarely find ourselves in such a favorable position. To quote a well-known example: Even the statement "If an animal is a bird, then it can fly." is not absolutely correct, since there are exceptions like penguins, ostriches etc. To deal with such cases—obviously, confining ourselves to absolutely correct conditionals would not be very helpful—we have to consider statistical syllogisms, statistical generalizations, and (particular) statistical explanations.

4.1 Statistical Syllogisms

By *statistical syllogism*, a term which was already mentioned in the introductory chapter, we mean a deductively shaped argument, in which the conditional is a statistical statement. For example, consider the following argument:

80% of the beans in box B are white.
This bean is from box B.

This bean is white.

(To make explicit the conditional form of the first premise, we may rewrite it as: "If x is a bean from box B, then x is white (with probability 0.8).")

The most important thing to note here is that with a statistical conditional a deductively shaped argument loses its distinctive mark, namely, it is no longer infallible. Since 20% of the beans in box B are not white, the conclusion may be false, even though the premisses are true. It would be a genuine deductive argument only, if the probability in the conditional were 1 (this is why we only called it a *deductively shaped* argument).

Nevertheless we are rather confident that the conclusion of the argument is true, since the probability of the bean being white is fairly high (provided the bean was picked at random). We may express this confidence by assigning to the conclusion a *degree of belief* (or a *degree of confidence*) equal to the probability stated in the conditional. We may even call this degree of belief a probability, if we switch from an *empirical* (or sometimes called *frequentistic*) interpretation of probabilities to a *subjective* (or *personalistic*) interpretation [Savage, 1954], or if we interpret the argument as a rule for decisions, which is to be applied several times. In the latter case the degree of belief measures the relative frequency of those cases in which a decision made according to this rule is correct (and thus can be seen as an empirical probability).[7]

[7] Note that we cannot interpret the number 0.8 assigned to the conclusion of the argument as an empirical probability for a single instance of the argument, since in a single instance the bean is either white or it is not white. There is no random element involved and thus the probability of the bean being white is either 0 or 1, even if we do not know whether it is 0 or 1.

4.2 Statistical Generalizations

A *statistical generalization*, which is called an inductive generalization in the introductory chapter, is an argument like, for instance:

If 50% of all children born are girls,
then 50% in this sample will be girls.
50% of the children in this sample are girls.
———————————————————
50% of all children born are girls.

Here the sample may be determined, for instance, as all children born in a specific hospital. (Note that often the first premise (the conditional) is missing, although the argument would not be logically complete without it.) Obviously, a statistical generalization is the probabilistic analog of an inductive argument. It would be a genuine inductive argument, if the percentages appearing in it were 100% (or the probabilities were 1). Statistical generalizations are one of the main topics of statistics, where they are used, for instance, to predict the outcome of an election from polls. In statistics it is studied what it takes to make reliable the conditional of the argument above (e.g. the sample has to be *representative* of the whole population), what is the best estimate of parameters for the whole population (like the relative frequency of 50% in the argument above—since it is, of course, logically possible that the relative frequency of girls in the whole population is 60%, but in the sample it is only 50%), and how to compute these best estimates.

4.3 (Particular) Statistical Explanations

We use *(particular) statistical explanation*[8] as a name for the probabilistic analog of an abductive inference. That is, a statistical explanation is an argument like

70% of all patients having caught a cold develop a cough.
Mrs Jones has a cough.
———————————————————
Mrs Jones has caught a cold.

Again, the above would be a genuine abductive argument, if the relative frequency in the conditional were 100% (the conditional probability were 1).

Although this argument seems plausible enough, there is a problem hidden in it, namely that we cannot assign a *degree of belief* to the conclusion as easily as we could for statistical syllogisms. The reason is that in a statistical syllogism the inference is drawn in the direction of the conditional probability—just as in a deductive argument the inference is drawn in the direction of the implication—whereas in a statistical explanation it is drawn in the opposite direction—just as in a reductive argument the inference is drawn opposite to the direction of the

[8] The word *particular* is added only to emphasize that the explanation must be a particular statement, since statistical generalizations—due to their reductive structure—also yield statistical explanations. In the following we drop it, since there is no danger of confusion.

implication. Thus, to assign a degree of belief to the conclusion of a statistical explanation, we have to invert the conditional probability using *Bayes' rule*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(B|A)$ is the conditional probability appearing in the first premise of the argument (with $A \hat{=}$ "The patient P has caught a cold." and $B \hat{=}$ "The patient P has a cough.") and $P(A)$ and $P(B)$ are the prior probabilities of the events A and B, respectively.[9]

It is obvious that using Bayes' rule we can always invert a conditional probability, provided the prior probabilities are known. That is, with Bayes' rule we can always transform a statistical explanation into a statistical syllogism, simply by turning the conditional around. Thus, in probabilistic reasoning the difference between abductively and deductively shaped arguments vanishes—which is not surprising, since the distinctive mark of deductive inferences, their infallibility, is lost (see above).

However, the distinction between arguments the conclusion of which is a general statement and those the conclusion of which is a particular statement remains valid, since it is next to impossible to know or even define the prior or posterior probability of a general statement. This is especially true, if the conclusions to be inferred are whole theories and consequently [Popper, 1934] has argued (convincingly, in our opinion) that there is no way to prefer, for instance, Einstein's theory of gravitation over Newton's based on their respective probabilities. Therefore in the following we confine ourselves to inferences of particular statements, which is no real restriction, since our topic is abductive inference anyway.

5 A GENERAL MODEL OF ABDUCTIVE INFERENCE

In this section we introduce a general model of abductive inference incorporating hypothesis assessment that is based on [Bylander *et al.*, 1991], but also closely related to [Peng and Reggia, 1989]. Since this model cannot be implemented directly (it would require too much storage space) we have to look for simplifications, which finally lead us to probabilistic networks.

5.1 Formal Definition

We start by giving a formal definition of an abductive problem. This definition is intended to describe the framework in which the abductive reasoning takes place by fixing the statements that may be used in abductive arguments and a mechanisms to assess the hypotheses resulting from such arguments.

[9] Note that Bayes rule is of little value for genuine abductive arguments, since for these prior and posterior probability must be the same (since they must be either 0 or 1) and thus $P(A|B) = P(A)$. That is, in order to apply Bayes' rule in this case, we already have to know what we want to infer.

DEFINITION 1. An **abductive problem** is a tuple $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$, where

- D_{all} is a finite set of possible atomic data,
- H_{all} is a finite set of possible atomic hypotheses,
- e is a relation of $2^{D_{\text{all}}}$ and $2^{H_{\text{all}}}$, i.e., $e \subseteq 2^{D_{\text{all}}} \times 2^{H_{\text{all}}}$,
- pl is a mapping from $2^{D_{\text{all}}} \times 2^{H_{\text{all}}}$ to a partially ordered set Q, and
- $D_{\text{obs}} \subseteq D_{\text{all}}$ is the set of observed data.

The sets D_{all} and H_{all} contain the statements that can be used in inferences, the former the possible observations (or *data*, which is why this set is denoted by the letter D), the latter the possible *hypotheses*. Since we deal with abductive inferences we require all statements in D_{all} and H_{all} to be particular, i.e., expressible as ground formulae. Note that we do not require the two sets to be disjoint, although this requirement is useful for most applications.

The relation e (for *explanation*) connects sets of observations with sets of hypotheses which explain them. That is, it describes the set of conditionals that can be used in abductive arguments (if $(H, D) \in e$, then $H \rightarrow D$). It thus contains information about what is accepted as an explanation (cf. section 3).

The mapping pl (for *plausibility*) assesses an inference by assigning an element of the set Q, which states the *quality* of the inferred (compound) hypothesis w.r.t. the given data. It may be a partial mapping, since we need values only for the elements of the relation e. In the following the set Q will always be the interval $[0, 1]$, as we deal only with probabilities and degrees of belief. However, this restriction is not necessary and therefore we chose to make the definition more general. Note that w.r.t. the mapping pl we may choose to add to D_{all} statements that do not need an explanation, but may have a bearing on the assessment of possible hypotheses. For example, in medical diagnosis we register the sex of a patient, although it is clearly not a fact to be explained by the diagnosis, simply because the likelihood of certain diseases differs considerably for the two sexes.

D_{obs} is the set of *observed* data, i.e., the set for which an explanation is desired. Of course, this set is not actually a part of the framework for abductive reasoning, which is fixed by the first four elements of the five-tuple. This framework may be used with several different sets D_{obs} and therefore it is reasonable to consider a general representation that is independent of D_{obs}. However, it has to be admitted that without a set of observations to be explained, there is no *problem* to be solved, and therefore, in contrast to [Bylander *et al.*, 1991], we chose to add it to the definition.

Before we can proceed further, we have to say a few words about the interpretation of subsets of the sets D_{all} and H_{all}, for instance, the interpretation of D_{obs}. Given a set $D \subseteq D_{\text{all}}$, we assume that *all* statements contained in D must hold, i.e., D is interpreted as a conjunction of its elements. This is straightforward, but what about the statements in $D_{\text{all}} \backslash D$? Obviously it would be impractical to require that they are all false, since this is equivalent to requiring perfect knowledge about all possible observations, which in applications is rarely to be had. Unfortunately, we can neither assume that nothing is known about their truth value, because there may be semantical relations between possible observations. For example, D_{all} may contain the statements “This bean is white.”, “This bean is red.” and “This bean is green.” As soon as we observe that the given bean is white, we know that the other two statements are false. (Obviously the same problem arises, if we consider whether a given set D of observations is satisfiable or whether it contains mutually exclusive statements.) Since this is a problem on the semantical level, it cannot be taken care of automatically in the formal model. Rather we have to assume that the relation e is defined in such a way that it connects only consistent sets of observations with consistent sets of hypotheses. In addition we have to assume that the set D_{obs} is consistent (which is no real restriction, since the observed reality should be consistent).

We can support that these requirements are met, though, by introducing some structure on the sets D_{all} and H_{all}. The simplest way to achieve this is to group atomic observations and hypotheses in such a manner that the statements in each group are mutually exclusive and exhaustive (to achieve the latter it may be necessary to introduce additional atomic statements that cover the remaining situations). After we did this, we can reconstruct the sets D_{all} and H_{all} as the Cartesian product of the domains of certain *variables*, each of which represents a group of mutually exclusive and exhaustive statements. That a set of observations and hypotheses is (formally) consistent can now easily be ensured by requiring that it must not assign more than one value to a variable, i.e., must not select more than one statement from the corresponding set of mutually exclusive statements. It should be noted, though, that this approach only excludes formal (or syntactical) inconsistencies, whereas factual inconsistencies (for instance, a nine year old girl with two children or a car that weighs 10 grams) still have to be taken care of by the explanation relation e.

An abductive problem is solved by finding the best explanation(s) for the data observed. This motivates the next two definitions.

DEFINITION 2. In an abductive problem $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ a set $H \subseteq H_{\text{all}}$ is called an **explanation** (of the data D_{obs}), iff $(H, D_{\text{obs}}) \in e$.

Often an explanation is required to be parsimonious (see above). We may add this requirement by defining

$$H \text{ is an explanation, iff } \ (H, D_{\text{obs}}) \in e \ \wedge \ \neg\exists H' \subset H : (H', D_{\text{obs}}) \in e,$$

that is, H is an explanation only, if no proper subset of H is also an explanation.

DEFINITION 3. In an abductive problem $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ an explanation H is called a **best explanation** (of the data D_{obs}), iff there is no explanation H' that is better than H w.r.t. the mapping pl, i.e., iff

$$\neg\exists H' : (H', D_{\text{obs}}) \in e \ \wedge \ \ pl(H, D_{\text{obs}}) < pl(H', D_{\text{obs}}).$$

Clearly, which explanation is selected depends on how the hypothesis assessment function pl ranks the possible explanations. Therefore this function should be chosen with special care. Fortunately, we can identify the ideal choice, which may serve as a guideline.

DEFINITION 4. The **optimal hypothesis assessment function** is

$$pl_{\mathrm{opt}}(H, D) = P(H|D).$$

The best explanation under pl_{opt} is called the **most probable explanation**.

Strictly speaking, the notation $P(H|D)$ used in this definition is formally incorrect, because the sets H and D are no events. However, this notation is intended as an abbreviation for $P\left(\left\{\omega \in \Omega \mid \bigwedge_{h \in H} h(\omega)\right\} \,\middle|\, \left\{\omega \in \Omega \mid \bigwedge_{d \in D} d(\omega)\right\}\right)$, where Ω is the underlying sample space. Here we view the statements h and d as random variables that assume the value *true*, if they hold for the elementary event ω, and the value *false* otherwise.

The probability of the hypothesis given the data is the optimal hypothesis assessment function, because it is easy to show that deciding on the hypothesis it advocates is, in the long run, superior to any other decision strategy—at least w.r.t. the relative number of times the correct decision is made. If the alternatives carry different costs in case of a wrong decision, an different function may be better. Nevertheless, the probability of the hypothesis given the data is still very important in this case, because it is needed to compute the hypothesis assessment function that optimizes the expected benefit.

Note that regarding a probabilistic assessment function as the best one possible does not exclude other uncertainty calculi. For a probabilistic approach to be feasible, specific conditions have to hold, which may not be satisfied in a given application. For example, if the available information about the modeled domain is not precise enough to compute a probability distribution, other approaches have to be considered. This justifies using, for example, possibility distributions [Dubois and Prade, 1988] or mass assignments [Baldwin *et al.*, 1995] for hypothesis assessment. In this chapter, however, we assume that we can treat abductive problems probabilistically and therefore drop the subscript *opt* in the following. An approach based on possibility distributions, which is closely related to the one presented here, can be found in the next chapter.

The relation e and the mapping pl of an abductive problem can easily be represented as a table in which the lines correspond to the possible sets of hypotheses and the columns correspond to the possible sets of observations (or vice versa, if you like). A sketch is shown in table 1. In this table each entry states the value assigned by the mapping pl to the pair (H, D) corresponding to the table field, provided that this pair (H, D) is contained in the relation e. For pairs (H, D) not contained in the relation e the value of the mapping pl is replaced by a 0, which serves as a kind of marker for unacceptable explanations. This way of marking explanations presupposes that any acceptable explanation has a non-vanishing probability, which is a reasonable assumption.

Table 1. Sketch of a possible representation of the relation e and the mapping pl of an abductive problem. The zeros indicate unacceptable explanations, i.e., pairs (H, D) not contained in e. The p_i are the conditional probabilities $P(H|D)$.

pl	$\emptyset$	$\{d_1\}$	$\{d_2\}$	$\{d_3\}$	$\cdots$	$\{d_1, d_2\}$	$\{d_1, d_3\}$	$\cdots$
$\emptyset$	1	0	0	0	$\cdots$	0	0	$\cdots$
$\{h_1\}$	0	p_1	p_6	p_{10}		p_{13}	p_{17}	
$\{h_2\}$	0	p_2	p_7	0		p_{14}	p_{18}	
$\{h_3\}$	0	p_3	0	p_{11}		0	p_{19}	
$\vdots$	$\vdots$				$\ddots$		$\vdots$	
$\{h_1, h_2\}$	0	p_4	p_8	0		p_{15}	0	
$\{h_1, h_3\}$	0	p_5	p_9	p_{12}	$\cdots$	p_{16}	p_{20}	$\cdots$
$\vdots$	$\vdots$						$\vdots$	$\ddots$

Note that inconsistent sets of observations or hypotheses correspond to all zero columns or lines and thus may be removed from the table. Note also that we actually need to mark the acceptable explanations, because we *cannot* define the relation e by

$$e = \left\{(H, D) \in 2^{H_{\text{all}}} \times 2^{D_{\text{all}}} \mid P(H|D) > 0\right\}.$$

We *cannot* even define the relation e as

$$e = \left\{(H, D) \in 2^{H_{\text{all}}} \times 2^{D_{\text{all}}} \mid P(H|D) > P(H)\right\}.$$

If we did so, we would accept unreasonable explanations. For example, if we learn that one or several bathing accidents have occurred, there is a much higher probability that the ice-cream sales have been high recently than without the bathing accidents. However, the high ice-cream sales are obviously not an acceptable explanation for the bathing accidents.[10] Generalizing, we cannot define the relation e as shown above, because what makes an explanation acceptable is a semantical relation between the hypotheses and the data. However, probability theory, just as logic, cannot capture semantical relations. Therefore we have to indicate separately which explanations are acceptable.

Solving an abductive problem with the representation just discussed is especially simple. All one has to do is to visit the table column that corresponds to the observed data and to find the line of this column that holds the highest probability. The set of hypotheses corresponding to this line is the most probable explanation for the observed data. However, it is clear that for any real world problem worth considering we cannot set up the table described above, since it would have too many lines and columns. Therefore we have to look for simplifications.

[10] The reason for this correlation is the simple fact that most bathing accidents occur in summer, because more people go bathing when it is warm. They also buy more ice-cream when it is warm.

5.2 Simplifications

In the following we consider, in two steps, simplifications of the general model introduced in the preceding section. The first simplification is based on the idea to replace the explanation relation e by a function mapping from $2^{H_{\text{all}}}$ to $2^{D_{\text{all}}}$, which assigns to a set H of hypotheses the union of all sets D of observations that H can explain. In this case we may just check for set inclusion to determine whether H explains D_{obs}. Of course, this simplification is not always possible, if we want the result to be equivalent to the original abductive problem. Specific conditions have to hold, which are given in the following definition.

DEFINITION 5. An abductive problem $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ is called **functional**, iff

1. $\forall H \subseteq H_{\text{all}} : \quad \forall D_1, D_2 \subseteq D_{\text{all}} :$
   $((H, D_1) \in e \wedge D_2 \subseteq D_1) \Rightarrow (H, D_2) \in e$

2. $\forall H \subseteq H_{\text{all}} : \quad \forall D_1, D_2 \subseteq D_{\text{all}} :$
   $((H, D_1) \in e \wedge (H, D_2) \in e \wedge D_1 \cup D_2$ is consistent$) \Rightarrow (H, D_1 \cup D_2) \in e$

Note that the first condition is no real restriction, since we interpret sets of observations as conjunctions. Consequently, if a set $D \subseteq D_{\text{all}}$ can be explained by a set $H \subseteq H_{\text{all}}$, all subsets of D should also be explainable by H. The second condition, however, is a restriction, since counterexamples can easily be found. For instance, due to the sexual dimorphism in mallards (*Anas platyrhynchos*), we can explain the observation of a female bird with webbings as well as the observation of a bird with webbings and a shining green-black head by the hypothesis that the bird is a mallard. However, the conjunction of the two observations cannot be explained by this hypothesis, since only male mallards have a shining green-black head [Thiele, 1997].[11]

The relation e of a functional abductive problem can be replaced, as the term "functional" already indicates, by a function e_f defined as follows

$$\forall H \subseteq H_{\text{all}} : \quad e_f(H) = \{d \in D_{\text{all}} \mid \exists D \subseteq D_{\text{all}} : d \in D \wedge (H, D) \in e\}.$$

With this function an explanation (for a consistent set D_{obs}) can be defined—as already indicated—as a set $H \subseteq H_{\text{all}}$, such that $D_{\text{obs}} \subseteq e_f(H)$.

The simplification that can be achieved for a functional abductive problem becomes most obvious, if the function e_f is represented as a relation e_1 of $2^{H_{\text{all}}}$ and D_{all}, which can be defined by

$$\forall H \subseteq H_{\text{all}} : \quad (H, d) \in e_1 \quad \Leftrightarrow \quad d \in e_f(H).$$

This relation can be represented as a table with one column for each $d \in D_{\text{all}}$ and one line for each $H \subseteq H_{\text{all}}$. Of course, lines corresponding to inconsistent sets

[11] The conjunction of the observations is not inconsistent, though, since in another kind of ducks, *tadorna tadorna,* which exhibits no sexual dimorphism w.r.t. plumage, male *and* female have a shining green-black head [Thiele, 1997].

Table 2. Sketch of a possible representation of the relation e and the mapping pl of a D-independent abductive problem. The zeros indicate possible observations that cannot be explained by the corresponding set of hypotheses.

pl	d_1	d_2	d_3	$\cdots$	d_n
$\emptyset$	p_1	p_2	p_3	$\cdots$	p_4
$\{h_1\}$	p_5	0	p_6		p_7
$\{h_2\}$	p_8	p_9	0		0
$\{h_3\}$	0	0	p_{10}		p_{11}
$\vdots$	$\vdots$				$\vdots$
$\{h_1, h_2\}$	p_{12}	p_{13}	p_{14}		p_{15}
$\{h_1, h_3\}$	p_{16}	0	p_{17}		p_{18}
$\vdots$	$\vdots$				$\vdots$

of hypotheses can be removed, just as for the table used to represent the original relation e. However, this table only represents the relation e in a more compact form. To state the hypothesis assessment function pl we still need a table with $2^{|H_{\text{all}}|}$ lines and $2^{|D_{\text{all}}|}$ columns. In order to represent pl with a table similar to the one needed to represent e_1, we need much stronger assumptions.

DEFINITION 6. A functional abductive problem $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ is called **D-independent**, iff

$$\begin{aligned} &\forall H \subseteq H_{\text{all}} : \forall D \subseteq D_{\text{all}} : \\ &\qquad D \text{ is consistent} \quad \Rightarrow \quad P(D|H) = \prod_{d \in D} P(d|H). \end{aligned}$$

Intuitively, D-independence means that the probability of a possible observation d is independent of any other observations that may be present given any set of hypotheses H. Note that the set H may be empty. That is, for all $D \subseteq D_{\text{all}}$ we also have $P(D) = \prod_{d \in D} P(d)$.

The relation e and the mapping pl of a D-independent abductive problem can easily be represented as a table in which the lines correspond to possible sets of hypotheses and the columns correspond to possible (atomic) observations. A sketch is shown in table 2. Each field of this table holds a conditional probability $P(d|H)$. Note that, as for table 1, we can delete from this table all lines that correspond to inconsistent sets of hypotheses. From this table we can then compute the assessment of any consistent set of hypotheses using Bayes' rule, provided we also know the prior probability $P(H)$ of the set of hypotheses:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = P(H) \prod_{d \in D} \frac{P(d|H)}{P(d)}.$$

Thus the best explanation is found by computing $P(H|D_{\text{obs}})$ for all (consistent) sets H of hypotheses using the above formula and then selecting that set H for which this probability is highest. This approach, although restricted to one element sets $\{h\}$, was suggested in [Charniak and McDermott, 1985].

As we have shown, the representation of the explanation relation e and the hypothesis assessment function pl of an abductive problem can be considerably simplified, if it is D-independent. However, the table of conditional probabilities required is still too large to be stored for any reasonably sized real world problem, simply because it still has too many lines. Therefore, in a second simplification step, we consider independent abductive problems. In such problems the possible hypotheses independently explain observations.

DEFINITION 7. A functional abductive problem $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ is called **independent**, iff

$$\forall H \subseteq H_{\text{all}} : \quad e_f(H) = \bigcup_{h \in H} e_f(\{h\}),$$

where e_f is the function by which the relation e can be represented in a functional abductive problem (see above).

Of course, this is a very strong restriction. It requires that there is no interaction of hypotheses in the sense that no combination of atomic hypotheses can to explain an observation, neither of the atomic hypotheses can explain on its own. In addition, the explanatory powers of atomic hypothesis must not cancel each other (there must not be "destructive interference"). This requirement excludes such commonplace situations as the following: Suppose we have a computing center that is equipped with its own emergency power supply. If we observe that the computing center is not working, we can explain this neither by the assumption that the electric power station is down nor by the assumption that the emergency power supply is broken. Only a conjunction of both is an acceptable hypothesis. On the other hand, cancellation must not occur, that is, one hypothesis must not have a subtractive effect on another. This is common in medicine, though. For example, in the domain of acid-base disorders, one disease might explain an increased blood pH, and another might explain a decreased pH, but together the result might be a normal pH ([Patil *et al.*, 1982] as cited in [Bylander *et al.*, 1991]).

Although excluding interactions between hypotheses is a strong restriction, we may choose to do so, because it enables us to simplify the representation of the framework for abductive reasoning. In an independent abductive problem the relation e can be replaced by a relation e_2, which is defined as follows:

$$\forall h \in H_{\text{all}} : \quad (h, d) \in e_2 \quad \Leftrightarrow \quad d \in e_f(h).$$

With this relation, an explanation can be defined as a consistent set H of hypotheses, such that

$$D_{\text{obs}} \subseteq \{d \in D_{\text{all}} \mid \exists h \in H : (h, d) \in e_2\}.$$

The relation e_2 can be represented as a simple table with $|H_{\mathrm{all}}|$ lines and $|D_{\mathrm{all}}|$ columns. In this table an entry corresponding to a pair (h, d) is marked, if the atomic hypothesis h can explain the atomic datum d. From this table all possible explanations H of the observed data D_{obs} can easily be constructed, if it is taken care that only consistent sets H are selected. However, as in the first simplification step, only the relation e can be represented in a more compact form. To store the hypothesis assessment function pl, we still need a table with $2^{|H_{\mathrm{all}}|}$ lines and $|D_{\mathrm{all}}|$ columns. In order to represent pl with a table similar to the one needed to represent e_2, we need much stronger assumptions [Borgelt, 1992]:

DEFINITION 8. A D-independent abductive problem $\mathcal{AP} = \langle D_{\mathrm{all}}, H_{\mathrm{all}}, e, pl, D_{\mathrm{obs}}\rangle$ is called **HD-independent**, iff

$$\begin{array}{l} \forall H \subseteq H_{\mathrm{all}} : \forall d \in D_{\mathrm{all}} : \\ H \text{ is consistent } \Rightarrow P(H|d) = \prod_{h \in H} P(h|d) \ \wedge \ P(H) = \prod_{h \in H} P(h). \end{array}$$

The idea underlying this definition is the following: To compute the assessment of a set H of hypotheses in a D-independent abductive problem we needed $P(H)$ and $P(d|H)$. These probabilities force us to store a table with $2^{|H_{\mathrm{all}}|}$ lines. However, with the above relations, we can compute these probabilities from a much smaller set of probabilities. For $P(H)$ this is obvious as the computation formula is given in the definition. For $P(d|H)$ we have to apply Bayes' rule:

$$P(d|H) = \frac{P(H|d)P(d)}{P(H)} = P(d) \prod_{h \in H} \frac{P(h|d)}{P(h)}.$$

Another application of Bayes' rule to the factors of the product of the above formula yields

$$P(d|H) = P(d) \prod_{h \in H} \frac{P(d|h)P(h)}{P(d)P(h)} = P(d) \prod_{h \in H} \frac{P(d|h)}{P(d)}.$$

Combining this formula and the one derived for D-independent abductive problems we arrive at the following formula to compute the conditional probability of a set H of hypotheses given a set D of possible observations [Borgelt, 1992]:

$$P(H|D) = \prod_{h' \in H} P(h') \prod_{d \in D} \left(\prod_{h \in H} \frac{P(d|h)}{P(d)} \right).$$

As a consequence we only need to store the probabilities $P(h)$, $P(d)$, and $P(d|h)$ for all $h \in H_{\mathrm{all}}$ and all $d \in D_{\mathrm{all}}$, that is, only $(|H_{\mathrm{all}}| + 1) \cdot (|D_{\mathrm{all}}| + 1) - 1$ probabilities. These probabilities can be organized in a simple table as sketched in table 3. In addition to this table we need the table that represents the relation e_2, because we cannot incorporate the relation e_2 in this table by marking entries with a zero (since we need all probabilities). Nevertheless, this representation is easily manageable even for large sets H_{all} and D_{all}.

Table 3. Sketch of a possible representation of the mapping pl of an HD-independent abductive problem.

pl	prior	d_1	d_2	$\cdots$	d_n
prior		p_{01}	p_{02}	$\cdots$	p_{0n}
h_1	p_{10}	p_{11}	p_{12}		p_{1n}
h_2	p_{20}	p_{21}	p_{22}		p_{2n}
$\vdots$	$\vdots$			$\ddots$	$\vdots$
h_m	p_{m0}	p_{m1}	p_{m2}	$\cdots$	p_{mn}

6 PROBABILISTIC NETWORKS

The simplifications introduced in the preceding section finally led to a manageable model of probabilistic abductive inference. However, the assumptions needed to arrive at this model are hardly acceptable. In practice we rarely encounter a problem in which all independence assumptions are satisfied. To cope with this problem, the idea suggests itself to search for a model, in which we can take dependences into account, but nevertheless can exploit all existing independences to reduce the amount of storage needed and to make inferences in complex domains tractable. Probabilistic networks are such a model.

6.1 Decomposition and Abductive Reasoning

Stated as concisely as possible, the basic ideas underlying probabilistic networks are these: Under certain conditions a probability distribution P on a multi-dimensional domain, which encodes *prior* or *generic knowledge* about this domain, can be decomposed into a set $\{P_1, \ldots, P_n\}$ of probability distributions on lower-dimensional subspaces. This decomposition is based on dependence and independence relations between the attributes of the domain. If such a decomposition is possible, it is sufficient to know the distributions on the subspaces to compute all probabilities that can be computed using the original distribution P. Since such a decomposition is usually represented as a network (or graph), it is commonly called a *probabilistic network* or a *probabilistic graphical model*. Reasoning in such a network consists in conditioning the represented probability distribution on the observed values of some attributes.

To elaborate a little: By *multi-dimensional domain* we mean that each state of the section of the world to be modeled can be described by stating the values of a finite set $V = \{A_1, \ldots, A_m\}$ of attributes. Each such attribute—or, more precisely, the set $\mathrm{dom}(A_k) = \{a_{k,1}, \ldots, a_{k,n_k}\}$ of its possible values—forms a dimension of the domain. Of course, to form a dimension, the possible values have to be *exhaustive* and *mutually exclusive*. With these restrictions the *universe of discourse* or *frame of discernment* is the multi-dimensional domain

$\Omega = \mathrm{dom}(A_1) \times \cdots \times \mathrm{dom}(A_m)$. Each possible state of this domain is described by a tuple $\omega = (a_{1,i_1}, \ldots, a_{m,i_m})$ containing the values which the attributes in V assume for this state.

The *probability distribution* P on Ω assigns to each point $\omega \in \Omega$ the (prior) probability that the modeled section of the world is in a state corresponding to that point. These probabilities are usually estimated by human domain experts or computed from a statistical analysis of available data.

By *decomposition* we mean that the probability distribution P on the domain as a whole can be reconstructed from the distributions $\{P_1, \ldots, P_r\}$ on lower-dimensional subspaces, where a subspace $\Omega_W \subseteq \Omega$ is the joint domain of a subset $W \subseteq V$ of attributes, i.e., $\Omega_W = \times_{A \in W} \mathrm{dom}(A)$. The distributions $\{P_1, \ldots, P_r\}$ on the subspaces assign conditional or marginal probabilities (depending on the network type—see below) to projections of the tuples $\omega \in \Omega$ to the corresponding subspaces. From these probabilities the original probability distribution P on Ω can be recomputed.

A decomposition of a probability distribution has several advantages, the most important being that it can usually be stored much more efficiently and with less redundancy than the original distribution. However, just being able to store a probability distribution more efficiently would not be of much use for reasoning tasks, were it not for the possibility to draw inferences in the underlying domain using only the distributions $\{P_1, \ldots, P_r\}$ on the subspaces without having to reconstruct the original distribution P. If we have obtained *evidential knowledge* about the current state ω_0 of the domain under consideration, which consists in observations of the values of some of the attributes, we can *condition* the represented probability distribution on the observed values by passing the conditioning information from subspace distribution to subspace distribution until all have been updated. This process is usually called *evidence propagation*.

Obviously, mapping our general model of abductive inference to probabilistic networks is—for the most part—very simple. When considering the interpretation of subsets H and D of hypotheses and possible observations above, we already mentioned that it is useful to form groups of mutually exclusive and exhaustive statements to support checking for the requirement that these subsets are consistent. These groups of statements correspond to the attributes considered in probabilistic networks. We only have to join the sets H_{all} and D_{all} to form one universe of discourse. (If we need the information where the attributes came from, we can keep it in attribute markers.) The hypothesis assessment function pl corresponds directly to the probability distribution P on the domain, since from this probability distribution we can compute the probability $P(H|D)$ for all sets H and D. Thus the decomposition can be used to simplify the representation of the hypothesis assessment function pl. The observed data D_{obs} corresponds to the evidential knowledge. The only element of an abductive problem for which there is no direct analog is the explanation relation e, which identifies the acceptable explanations.

In the following sections we first review the theory of probabilistic networks in more detail and later we turn to the problem of hypothesis selection.

6.2 Conditional Independence

Whether and how a given probability distribution P can be be decomposed into a set $\{P_1, \ldots, P_r\}$ of distributions on subspaces is determined by the dependence structure of the attributes of the domain Ω underlying P. It is clear that attributes that directly depend on each other must appear together at least once in a distribution of the decomposition. A separation is possible only, if they are *conditionally independent* [Dawid, 1979; Pearl, 1992]. That is, if two attributes get independent, if certain other attributes are fixed, their dependence is not genuine, but mediated through other attributes. In this case these two attributes need not appear together in a distribution of the decomposition, but it suffices that there is a “path” connecting them through the mediating attributes. (This already suggests that a decomposition may be representable as a network or graph.)

Conditional probabilistic independence is defined in accordance with the usual notion of stochastic independence as follows:

DEFINITION 9. Let X, Y, and Z be three disjoint subsets of attributes in V. X is called **conditionally independent** of Y given Z w.r.t. P, written $X \perp\!\!\!\perp Y \mid Z$, iff

$$\forall \omega \in \Omega: \quad P(\omega_{X \cup Y} \mid \omega_Z) = P(\omega_X \mid \omega_Z) \cdot P(\omega_Y \mid \omega_Z)$$

whenever $P(\omega_Z) > 0$.

Here a conditional probability distribution is defined in the usual way, i.e., as

$$P(\omega_X \mid \omega_Y) = \frac{P(\omega_{X \cup Y})}{P(\omega_Y)}.$$

The notion of conditional independence provides, as already mentioned, the connection to a graph representation. It has been shown in general that a notion of conditional independence satisfying certain axioms, which are known as the *semi-graphoid axioms* [Dawid, 1979; Spohn, 1980; Pearl and Paz, 1987; Smith, 1989], can be used to define a graph structure on the set of attributes. These axioms are

symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$

decomposition: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z)$

weak union: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z \cup W)$

contraction: $(W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

The *symmetry* axiom states that in any state of knowledge Z, if Y tells us nothing new about X, then X tells us nothing new about Y. The *decomposition* axiom asserts that if two combined items of information are judged irrelevant to X, then each separate item is irrelevant as well. The *weak union* axiom states that learning irrelevant information W cannot help the irrelevant information Y become relevant to X. The *contraction* axiom states that if we judge W irrelevant to X after learning some irrelevant information Y, then W must have been irrelevant before we learned Y. Together the weak union and contraction properties mean that irrelevant information should not alter the relevance of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant [Pearl, 1992]. It is plausible that any reasonable notion of conditional independence should satisfy these axioms and, indeed, probabilistic conditional independence does.

6.3 Graph Representation

Given an appropriate notion of conditional independence, i.e., one that satisfies the semi-graphoid axioms, a *conditional independence graph* $G = (V, E)$ can be defined. Each node of this graph corresponds to an attribute of the underlying domain. The topology of the graph, i.e., which edges are present and which are missing, is used to represent an independence model, i.e., a set of conditional independence statements, of the domain under consideration [Pearl, 1992; Spirtes *et al.*, 1993]: The conditional independence of two sets of attributes given a third is expressed by *separation* of the corresponding node sets by the nodes that correspond to the conditioning attributes.

What is to be understood by "separation" depends on whether the graph is chosen to be directed or undirected. If it is undirected, separation is defined as follows: If X, Y, and Z are three disjoint subsets of nodes in an undirected graph, then Z separates X from Y, iff after removing the nodes in Z and their associated edges from the graph there is no path, i.e., no sequence of consecutive edges, from a node in X to a node in Y. Or, in other words, Z separates X from Y, iff all paths from a node in X to a node in Y contain a node in Z.

For directed graphs, which have to be acyclic, the so-called d*-separation criterion* is used [Pearl, 1992; Verma and Pearl, 1990]: If X, Y, and Z are three disjoint subsets of nodes in a directed acyclic graph (DAG), then Z is said to d*-separate* X from Y, iff there is no path, i.e., no sequence of consecutive edges (of any directionality), from a node in X to a node in Y along which the following two conditions hold:

1. every node with converging edges either is in Z or has a descendant in Z,
2. every other node is not in Z.

With the described notions of separation, we can define the so-called *Markov properties* of graphs [Whittaker, 1990]:

pairwise: Attributes, whose nodes are non-adjacent in the graph, are conditionally independent given all remaining attributes.

local: Given the attributes of the adjacent nodes (the neighbors), an attribute is conditionally independent of all remaining attributes.

global: Any two subsets of attributes, whose corresponding node sets are separated by a third node set, are conditionally independent given the attributes corresponding to the nodes in the third set.

Note that the local Markov property is contained in the global, and the pairwise Markov property in the local.

Since the pairwise Markov property refers to the independence of only two attributes, it would be most natural (at least for undirected graphs) to use it to define an independence graph: If two attributes are dependent given all other attributes, there is an edge between their corresponding nodes, otherwise there is no edge [Whittaker, 1990]. But, unfortunately, the three types of Markov properties are not equivalent in general, and it is obvious that we need the *global* Markov property for inferences from multiple observations. However, the above definition can be used, if—in addition to the semi-graphoid axioms—the following axiom holds:

intersection: $(W \perp\!\!\!\perp Y \mid Z \cup X) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

The semi-graphoid axioms together with this one are called the *graphoid axioms*. If they hold for a given notion of conditional independence, an independence graph can be defined via the pairwise Markov condition, since the intersection axiom allows us to infer the global Markov property from the pairwise. If the intersection axiom does not hold, the global Markov property has to be used to define an independence graph. If can be shown that a strictly positive probability distribution satisfies the intersection axiom [Pearl, 1992] and therefore the probability distribution on the domain to be modeled is often required to be strictly positive.

It is clear that an independence graph for a given domain is easy to find. For example, the complete undirected graph, i.e., the graph in which every node is connected directly to every other, always is an independence graph. Similarly, a directed acyclic graph which becomes a complete undirected graph if the directions of the edges are neglected is a trivial independence graph. However, using such graphs would not reduce the amount of data that needs to be stored (see below). Therefore we have to add the condition that the independence graph has to be *sparse* or even *minimal*, i.e., should contain as few edges as possible.

Whether directed acyclic graphs or undirected graphs are used is, to a certain degree, a matter of taste. However, it should be noted that the two types of graphs represent conditional independence relations in fundamentally different ways. There are undirected graphs that represent a set of conditional independence statements that cannot be represented by a single directed acyclic graph, and vice versa.

6.4 Factorization

The conditional independence graph is also called the *qualitative* part of a probabilistic network, since it specifies which attributes are dependent and which are independent, but not the details of the dependences. These are represented in the *quantitative* part of a probabilistic network which consists of a set of probability distributions. W.r.t. the quantitative part of a probabilistic network the conditional independence graph describes a *factorization* of the joint probability distribution P on Ω. The exact representation of the quantitative information and the factorization formula depends on the type of the conditional independence graph.

Bayesian networks. The most popular kind of probabilistic networks in artificial intelligence is the *Bayesian network*, also called *belief network* [Pearl, 1986; Pearl, 1992]. A Bayesian network consists of a directed acyclic graph and a set of conditional probability distributions $P(\omega_A \mid \omega_{\mathrm{parents}(A)})$, $A \in V$, where $\mathrm{parents}(A)$ is the set of attributes corresponding to the parents of the node that corresponds to attribute A. That is, there is one probability distribution for each attribute and each distinct instantiation of its parent attributes in the graph. If an attribute does not have any parents, its associated distribution is simply an unconditional distribution.

A Bayesian network describes a factorization of a strictly positive joint probability distribution P on Ω w.r.t. a directed acyclic graph into a set of conditional probability distributions according to

$$\forall \omega \in \Omega : \quad P(\omega) = \prod_{A \in V} P(\omega_A \mid \omega_{\mathrm{parents}(A)}).$$

These equations can easily be derived from the (generally valid) *chain rule of probability*,

$$\forall \omega \in \Omega : \quad P(\omega) = \prod_{i=1}^{m} P(\omega_{A_i} \mid \omega_{\{A_1,\ldots,A_{i-1}\}}),$$

and a set of conditional independence statements. The reason is that a conditional independence statement $X \perp\!\!\!\perp Y | Z$ implies

$$P(\omega_X \mid \omega_{Y \cup Z}) = P(\omega_X \mid \omega_Z).$$

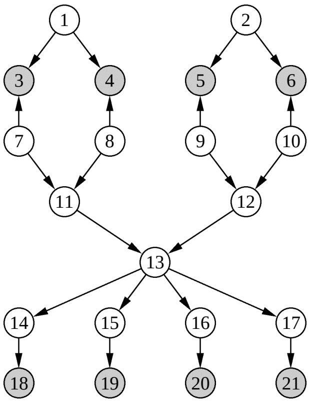
Therefore we can cancel from the conditions all attributes of which the attribute A_i is independent given the remaining attributes and thus arrive at the factorization formula shown above. It should be noted, though, that the achievable simplification may depend on the order of the attributes.

Markov networks. An alternative type of probabilistic networks uses undirected graphs and is called a *Markov network* [Lauritzen and Spiegelhalter, 1988; Pearl, 1992]. It represents so-called Markov random fields. Similar to a Bayesian network a Markov network describes a factorization of the joint probability distribution P on Ω, but it uses a *potential representation*: A strictly positive probability distribution P on Ω factorizes w.r.t. an undirected graph $G = (V, E)$, iff

$$\forall X \in \mathrm{cliques}(G) : \exists \phi_X : \forall \omega \in \Omega : \quad P(\omega) = \prod_{X \in \mathrm{cliques}(G)} \phi_X(\omega_X),$$

where $\mathrm{cliques}(G)$ is the set of all maximal cliques of G, each of which is represented by the set of attributes whose corresponding nodes are contained in it. The *factor potentials* ϕ_X are strictly positive functions defined on Ω_X, $X \subseteq V$.

Whether a Bayesian network or a Markov network is chosen to model a given domain is, as already indicated above, to a certain degree a matter of taste. However,



21 attributes:
1 – dam correct?
2 – sire correct?
3 – stated dam ph.gr. 1
4 – stated dam ph.gr. 2
5 – stated sire ph.gr. 1
6 – stated sire ph.gr. 2
7 – true dam ph.gr. 1
8 – true dam ph.gr. 2
9 – true sire ph.gr. 1
10 – true sire ph.gr. 2
11 – offspring ph.gr. 1
12 – offspring ph.gr. 2
13 – offspring genotype
14 – factor 40
15 – factor 41
16 – factor 42
17 – factor 43
18 – lysis 40
19 – lysis 41
20 – lysis 42
21 – lysis 43

The grey nodes correspond to observable attributes.

Figure 2. Conditional independence graph of a graphical model for genotype determination and parentage verification of Danish Jersey cattle in the F-blood group system [Rasmussen, 1992].

since a Bayesian network is a directed graph, it is well-suited to represent direct causal dependencies between variables: Often we can choose the directions of the edges (and thus the “directions” of the conditional probabilities) in such a way that they coincide with the direction of the causal influence. This is quite natural for knowledge representation, especially in applications for diagnostic reasoning, i.e., abductive inference, and thus one may prefer Bayesian network models. However, the causal interpretation of Bayesian networks should be handled with care, since it involves strong assumptions about the statistical manifestation of causal dependences [Borgelt and Kruse, 1999].

6.5 An Example Network

As an example of a probabilistic network we consider an application for blood group determination of Danish Jersey cattle in the F blood group system, the primary purpose of which is parentage verification for pedigree registration [Rasmussen, 1992]. The underlying domain is described by 21 attributes, eight of which are observable. The size of the domains of these attributes ranges from two to eight possible values and the total frame of discernment has $2^6 \cdot 3^{10} \cdot 6 \cdot 8^4 = 92\,876\,046\,336$ possible states. This space is clearly much too large to be handled as a whole and therefore a decomposition of the expert knowledge about this domain is necessary to make reasoning feasible. Figure 2 lists the attributes and shows the conditional independence graph of this graphical model, which was designed by human domain experts (the graphical model is a Bayesian network and thus the conditional independence graph is a directed acyclic graph). The grey nodes correspond to the observable attributes.

Table 4. Conditional probability distributions for a subgraph of the conditional independence graph shown in Figure 2.

sire	phenogroup	phenogroup stated sire		
correct	true sire	F1	V1	V2
yes	F1	1	0	0
yes	V1	0	1	0
yes	V2	0	0	1
no	F1	0.58	0.10	0.32
no	V1	0.58	0.10	0.32
no	V2	0.58	0.10	0.32

In this example, a decomposition of the joint probability distribution according to the conditional independence graph shown in figure 2 leads to a considerable simplification. Instead of having to determine the probability of each of the 92 876 046 336 elements of the 21-dimensional frame of discernment Ω, only 306 conditional probabilities in subspaces of at most three dimensions need to be specified. An example of a conditional probability table is shown in table 4, which is for the phenogroup 1 of the stated sire of a given calf conditioned on the phenogroup of the true sire of the calf and whether the sire was correctly identified. The numbers in this table are derived from statistical data and the experience of domain experts. The family of all 21 conditional probability tables forms the quantitative part of the graphical model for the Danish Jersey cattle example.

6.6 Evidence Propagation

After a probabilistic network has been constructed, it can be used to do reasoning. In the Danish Jersey cattle example, for instance, the phenogroups of the stated dam and the stated sire can be determined and the lysis values of the calf can be measured. With this information it becomes possible to assess whether the stated parents of the calf are the true parents or whether the breeder has made a mistake (or has attempted to deceive the registration office).

However, reasoning in a probabilistic network is not always completely straightforward. Considerations of efficiency make it often advisable to transform a graphical model into a form that is better suited for propagating the evidential knowledge and computing the resulting marginal distributions for the unobserved attributes. We briefly sketch here a popular efficient reasoning method known as *clique tree propagation* (CTP) [Lauritzen and Spiegelhalter, 1988; Castillo *et al.*, 1997], which involves transforming the conditional independence graph into a clique tree. A well-known interactive software tool for probabilistic reasoning in clique trees is HUGIN [Andersen *et al.*, 1989].

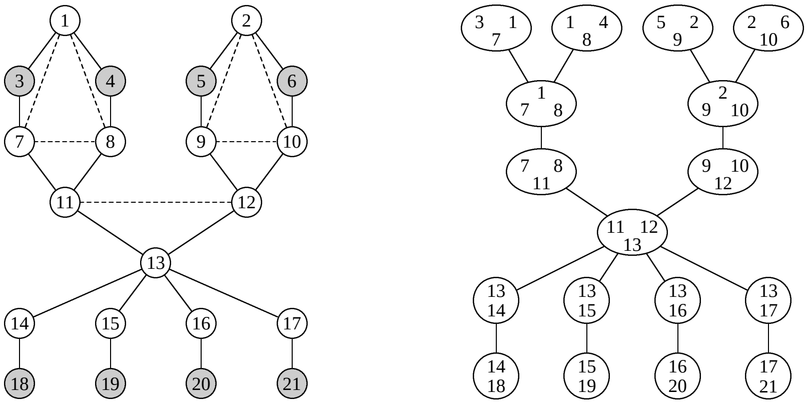


Figure 3. Triangulated moral graph (left) and clique tree (right) for the graphical model shown in Figure 2. The dotted lines are the edges added when parents were "married".

The transformation into a clique tree is carried out as follows: If the conditional independence graph is a directed acyclic graph, it is first turned into an undirected graph by constructing its associated *moral graph* [Lauritzen and Spiegelhalter, 1988]. A moral graph is constructed from a directed acyclic graph by "marrying" the parent nodes of all nodes (hence the name "moral graph"). This is done by simply adding undirected edges between the parents. The directions of all other edges are discarded. In general the moral graph satisfies only a subset of the independence relations of the underlying directed acyclic graph, so that this transformation may result in a loss of independence information. The moral graph for the Danish Jersey Cattle example is shown on the left in figure 3. The edges that were added when parents were "married" are indicated by dotted lines.

In a second step, the undirected graph is triangulated. (If the conditional independence graph is an undirected graph right from the start, this is the first step to be carried out.) An undirected graph is called *triangulated*, if all cycles containing at least four nodes have a chord, where a chord is an edge that connects two non-adjacent nodes of the cycle. To achieve triangulation, it may be necessary to add edges, which may result in a (further) loss of independence information. In the Danish Jersey cattle example, however, the moral graph shown on the left in figure 3 is already triangulated, so no new edges need to be introduced.

Finally, the triangulated graph is turned into a clique tree by finding the maximal cliques, where a clique (see above) is a fully connected subgraph, and it is maximal, if it is not contained in another clique. In the clique tree there is one node for each maximal clique of the triangulated graph and its edges connect nodes that represent cliques having attributes in common. It should be noted that in general the clique tree is not unique, because often different sets of edges can be chosen.

The clique tree for the Danish Jersey cattle example is shown on the right in figure 3. Detailed information on triangulation, clique tree construction and other related graph-theoretical problems can be found in [Castillo *et al.*, 1997].

The quantitative part of a graphical model, of course, has to be transformed, too. From the quantitative information of the original graphical model one has to compute a marginal distribution for each of the subspaces represented by the nodes of the clique tree. For the Danish Jersey cattle example, we have to compute a marginal distribution for the subspace formed by the attributes 1, 3, 7, one for the subspace formed by the attributes 1, 4, 8, and so on.

Having constructed a clique tree, which is merely a preliminary operation to make evidence propagation more efficient, we can finally turn to evidence propagation itself. Evidence propagation in clique trees is basically an iterative extension and projection process. When evidence about the value of an attribute becomes available, it is first extended to a clique tree node the attribute is contained in. This is done by conditioning the associated marginal distribution. We call this an extension, since by this conditioning process we go from restrictions on the values of a single attribute to restrictions on tuples of attribute values. Hence the information is extended from a single attribute to a subspace formed by several attributes. Then the conditioned distribution is projected to all intersections of the clique tree node with other nodes. Via these projections the information can be transferred to other nodes, where the process repeats: First it is extended to the subspace represented by the node, then it is projected to the intersections connecting it to other nodes. The process stops when all nodes have been updated.

The propagation scheme outlined above and the subsequent computation of posterior marginal distributions for the unobserved attributes can easily be implemented by locally communicating node- and edge-processors. These processor also serve the task to let pieces of information "pass" each other without interaction. Such bypassing is necessary, if the propagation operations in the underlying uncertainty calculus are not idempotent, that is, if incorporating the same information twice can invalidate the results. This is the case in probabilistic reasoning. This problem is also the reason why the clique graph is usually required to be a *tree*: If there were loops, information could travel on two or more different paths to the same destination and thus be incorporated twice.

Of course, clique tree propagation is not the only known evidence propagation scheme. Others include bucket elimination [Dechter, 1996; Zhang and Poole, 1996] and iterative proportional fitting [Whittaker, 1990]. Commonly used evidence propagation algorithms differ from each other w.r.t. the network structures they support. For example, both bucket elimination and iterative proportional fitting can also be used with networks that contain cycles.

6.7 Hypothesis Selection

As already indicated at the beginning of this section, probabilistic networks mainly provide means to represent the hypothesis assessment function pl. However, in a probabilistic network there is no direct analog to the explanation relation e of our general model of an abductive problem. That is, with a probabilistic network, we can easily compute the plausibility of a given hypothesis, but we cannot check whether the hypothesis is semantically acceptable.

In applications tailored for a specific application this is often irrelevant, because from the application it can be clear what attributes we are interested in and therefore we only have to compute the most probable tuple of values for the subspace formed by these attributes. As an example consider again the Danish Jersey Cattle example discussed above. Its primary purpose is, as already mentioned, parentage verification. That is, we are interested in the values of two variables, namely "Dam correct?" and "Sire correct?". All other unobservable variables are only there to structure the available expert knowledge about the domain. For the two attributes, however, which indicate whether the stated dam is the true dam and whether the stated sire is the true sire, we compute the probability of all possible combinations and select the most probable one. Actually, in the original network there is an additional node "Parent correct?" which combines these hypotheses in a single attribute. In the discussion above, we discarded this node, because it does not carry real information, but only makes the clique tree unnecessarily complex. In cases like these, where a set of attributes can be fixed in advance, a probabilistic network approach to abductive reasoning is most powerful.

To identify the most probable tuple in a subspace formed by a set of attributes, may not always be appropriate, though. If, for instance, not all acceptable (compound) hypotheses consist of the same number of atomic hypotheses, we cannot use it, because it may result in hypotheses too specific for a given problem. However, even in this case the probabilistic network alone may contain enough information to select the best acceptable hypothesis. For example, the structure of the network can provide information how to restrict the set of attributes we have to take into account to form a (compound) hypothesis. Obviously, it is sufficient to select a set of explanatory attributes (i.e., attributes derived from $H_{\rm all}$) in such a way that the observed attributes and the remaining attributes are conditionally independent given the selected attributes. A (compound) hypothesis formed from these attributes has to be considered complete because due to the interpretation of the semi-graphoid axioms (see above) the remaining attributes are irrelevant for the observations. Of course, such a restriction does not guarantee that the selected (compound) hypothesis is semantically acceptable, but it may help to restrict the set of hypotheses one has to consider. The idea can be enhanced by the requirement that a reasonable hypothesis should make the observed data more likely than it is without it or that the observed data should make an acceptable hypothesis more likely. However, as we showed with a simple example above, a higher likelihood is not sufficient for a semantically acceptable hypothesis.

If these approaches, which try to do without additional information, are not feasible, we have to add some structure to represent (a simplification of) the explanation relation e. A very manageable structure results, if we have an independent abductive problem and thus can represent the relation e as a simple table with

H_{all} lines and D_{all} columns. In this case the probabilistic network helps to avoid the strong probabilistic independence assumptions underlying D-independent and HD-independent abductive problems. We only need the logical independence assumptions needed to simplify the representation of the relation e. If even these logical assumptions are too strong, we can enhance the table of the relation e for an independent abductive problem by an explicit list of (compound) hypotheses, for which "constructive" or "destructive inference" occurs, i.e., those (compound) hypotheses which can explain more than the sum of their elements and those, which can explain less (examples of such situation we gave above). Provided this list is of moderate size, the problem remains tractable. A more detailed discussion can be found in [Bylander *et al.*, 1991; Borgelt, 1992].

6.8 Learning from Data

A probabilistic network is a powerful tool to support reasoning—as soon as it is constructed. Its construction by human experts, however, can be tedious and time consuming. Therefore recent research in probabilistic graphical models focused on learning them from a database of sample cases. In accordance with the two components of graphical models, one distinguishes between *quantitative* and *qualitative* (or *structural*) *network induction*.

Quantitative network induction for a given network structure consists in estimating the joint probability distribution P, where P is selected from a family of parameterized probability distributions. A lot of approaches have been developed in this field, using methods such as maximum likelihood, maximum penalized likelihood, or fully Bayesian approaches, which involve different computational techniques of probabilistic inference such as the expectation maximization (EM) algorithm, Gibbs sampling, Laplace approximation, and Monte Carlo methods. For an overview, see [Buntine, 1994; Spiegelhalter *et al.*, 1993].

Qualitative network induction consists in learning a network structure from a database of sample cases. In principle one could use the factorization property of a probabilistic network to evaluate its quality by comparing for each $\omega \in \Omega$ the probability computed from the network with the relative frequency found in the database to learn from. However, this approach is usually much too costly.

Other methods are based on linearity and normality assumptions [Pearl and Wermuth, 1993], rely on the extensive testing of conditional independences (CI tests) [Verma and Pearl, 1992], or use a Bayesian approach [Cooper and Herskovits, 1992; Lauritzen *et al.*, 1993]. Unfortunately, the first group is fairly restrictive, CI tests tend to be unreliable unless the volume of data is enormous, and with an increasing number of vertices they soon become computationally intractable. Bayesian learning requires debatable prior assumptions (for example, default uniform priors on distributions, uniform priors on the possible graphs) and also tends to be inefficient unless greedy search methods are used.

Nevertheless, several network induction algorithms have successfully been applied. The oldest example is an algorithm to decompose a multi-variate probability distribution into a tree of two-dimensional distributions [Chow and Liu, 1968]. It uses mutual information as an evaluation measure and optimum weight spanning tree construction as a search method.

Another example is the $K2$ algorithm [Cooper and Herskovits, 1992], which uses a greedy parent search and a Bayesian evaluation measure. Its drawback, which consists in the fact that it needs a topological order of the attributes, can be overcome by a hybrid algorithm [Singh and Valtorta, 1993], which combines CI tests (to find a topological order) and $K2$ (to construct the Bayesian network with respect to this topological order). Unfortunately, $K2$ can deal only with complete and precise data. The treatment of missing values and hidden variables is clear only from a theoretical point of view [Cooper and Herskovits, 1992].

A third algorithm, which uses a backward search strategy, has been described in [Højsgaard and Thiesson, 1994].

Several evaluation measures, which can be used with optimum weight spanning tree construction and greedy parent search as well as other search methods, are surveyed in [Borgelt and Kruse, 1997].

7 SUMMARY

In this paper we considered how probabilistic networks can support abductive reasoning. Starting from a definition of an abductive inference as a reductive, i.e., explanatory inference whose conclusion is a particular statement, we showed how probability theory enters ours consideration due to two reasons: In the first place, if we want to handle real world problems, we have to take into account statistical conditionals. Secondly, in order to reduce the chances of an incorrect result, we have to assess and compare the conclusions of abductive inferences. Based on a general model of abductive inference we showed that a direct approach to represent a hypothesis assessment function is not feasible and thus simplifications are required. Although straightforward simplifications lead to a manageable model, they involve strong presuppositions which cannot reasonably be expected to hold in applications. As a solution probabilistic networks suggest themselves as well-established technique to decompose a multi-variate probability distribution in order to make reasoning in high-dimensional domains possible. They are very well-suited to represent the hypothesis assessment function of abductive problem solving. However, it may be necessary to enhance them by a method to identify the acceptable hypotheses, because the raw probabilistic information they represent is often not sufficient for this task.

Department of Knowledge Processing and Language Engineering, Otto-von-Guericke-University of Magdeburg, Germany.

REFERENCES

[Andersen *et al.*, 1989] S. K. Andersen, K. G. Olesen, F. V. Jensen and F. Jensen. HUGIN — a shell for building Bayesian belief universes for expert systems. *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI'89, Detroit, MI)*, 1080–1085. Morgan Kaufman, San Mateo, CA, 1989.

[Baldwin *et al.*, 1995] J. F. Baldwin, T. P. Martin and B.W. Pilsworth. *FRIL — Fuzzy and Evidential Reasoning in Artificial Intelligence*. Research Studies Press/J. Wiley & Sons, Taunton/Chichester, 1995.

[Bochenski, 1954] I. M. Bocheński. *Die zeitgenössischen Denkmethoden*. Franke-Verlag, Tübingen, 1954.

[Borgelt, 1992] C. Borgelt. *Konzeptioneller Vergleich verschiedener numerischer und logischer Ansätze abduktiver Inferenz*. Diplomarbeit, TU Braunschweig, 1992.

[Borgelt and Kruse, 1997] C. Borgelt and R. Kruse. Evaluation measures for learning probabilistic and possibilistic networks. *Proceedings of the 6th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'97, Barcelona)*, Vol. 2:1034–1038. IEEE Press, Piscataway, NJ, 1997.

[Borgelt and Kruse, 1999] C. Borgelt and R. Kruse. A Critique of Inductive Causation. *Proc. 5th European Conf. on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'99, London), LNAI 1638*, 68–79. Springer, Heidelberg, 1999.

[Buntine, 1994] W. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research* 2:159–225, 1994.

[Bylander *et al.*, 1991] T. Bylander, D. Allemang, M. C. Tanner and J. R. Josephson. The computational complexity of abduction. *Artificial Intelligence* 49: 25–60. North-Holland, Amsterdam, 1991.

[Castillo *et al.*, 1997] E. Castillo, J. M. Gutierrez and A. S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer, New York, NY, 1997.

[Charniak and McDermott, 1985] E. Charniak and D. McDermott. *Introduction to Artificial Intelligence*. Addison-Wesley, Reading, MA, 1985.

[Chow and Liu, 1968] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Information Theory* 14(3):462–467. IEEE Press, Piscataway, NJ, 1968.

[Cooper and Herskovits, 1992] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9:309–347. Kluwer, Dordrecht, 1992.

[Dawid, 1979] A. Dawid. Conditional independence in statistical theory. *SIAM Journal on Computing* 41:1–31, 1979.

[Dechter, 1996] R. Dechter. Bucket elimination: a unifying framework for probabilistic inference. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI'96, Portland, OR)*, 211–219. Morgan Kaufman, San Mateo, CA, 1996.

[Dubois and Prade, 1988] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, NY, 1988.

[Hempel, 1966] C. G. Hempel. *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

[Højsgaard and Thiesson, 1994] S. Højsgaard and B. Thiesson. BIFROST — block recursive models induced from relevant knowledge, observations, and statistical techniques. *Computational Statistics and Data Analysis*, 1994.

[Josephson and Josephson, 1996] J. R. Josephson and S. G. Josephson. *Abductive Inference — Computation, Philosophy, Technology*. Cambridge University Press, Cambridge, MA, 1996.

[Kruse *et al.*, 1991] R. Kruse, E. Schwecke, and J. Heinsohn. *Uncertainty and Vagueness in Knowledge-based Systems: Numerical Methods (Series: Artificial Intelligence)*. Springer, Berlin, Germany 1991

[Lauritzen and Spiegelhalter, 1988] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 2(50):157–224. Blackwell, Oxford, 1988.

[Lauritzen *et al.*, 1993] S. L. Lauritzen, B. Thiesson and D. Spiegelhalter. Diagnostic systems created by model selection methods — a case study. *Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics (Fort Lauderdale, FL)*, pp. 93–105, 1993.

[Losee, 1993] J. Losee. *A Historical Introduction to the Philosophy of Science (3rd edition)*. Oxford University Press, Oxford, 1993.

[Patil *et al.*, 1982] R. S. Patil, P. Szolovits and W. B. Schwartz. Modeling knowledge of the patient in acid-base and electrolyte disorders. In: P. Szolovits, ed. *Artifical Intelligence in Medicine*, pp. 191–226. Westview Press, Boulder, CO, 1982.

[Pearl, 1986] J. Pearl. Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence* 29:241–288, 1986.

[Pearl, 1992] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition).* Morgan Kaufman, San Mateo, CA, 1992.

[Pearl and Paz, 1987] J. Pearl and A. Paz. Graphoids: A Graph Based Logic for Reasoning about Relevance Relations. In: B. D. Boulay et al., eds. *Advances in Artificial Intelligence 2,* pp. 357–363. North Holland, Amsterdam, 1987.

[Pearl and Wermuth, 1993] J. Pearl and N. Wermuth. When can association graphs admit a causal interpretation? *Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics (Fort Lauderdale, FL)*, pp. 141–150, 1993.

[Peirce, 1958] C. S. Peirce. (C. Hartshorne, P. Weiss and A. Burks, eds.) *Collected Papers of Charles Sanders Peirce*. Havard University Press, Cambridge, MA, 1958.

[Peng and Reggia, 1989] Y. Peng and J. A. Reggia. *Abductive Inference Models for Diagnostic Problem Solving.* Springer, New York, NY, 1989.

[Popper, 1934] K. R. Popper. *Logik der Forschung*. 1st edition: Julius Springer, Vienna, 1934. 9th edition: J. C. B. Mohr, Tübingen, 1989. English edition: *The Logic of Scientific Discovery,* Hutchinson, London, 1959.

[Rasmussen, 1992] L. K. Rasmussen. *Blood Group Determination of Danish Jersey Cattle in the F-blood Group System (Dina Research Report 8).* Dina Foulum, Tjele, 1992.

[Salmon, 1973] W. C. Salmon. *Logic (2nd edition).* Prentice-Hall, Englewood Cliffs, NJ, 1973.

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

[Singh and Valtorta, 1993] M. Singh and M. Valtorta. An algorithm for the construction of Bayesian network structures from data. *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, pp. 259–265. Morgan Kaufman, San Mateo, CA, 1993.

[Smith, 1989] J. Q. Smith. Influence Diagrams for Statistical Modeling. *Annals of Statistics* 17(2):654–672, 1989.

[Spiegelhalter *et al.*, 1993] D. Spiegelhalter, A. Dawid, S. Lauritzen and R. Cowell. Bayesian Analysis in Expert Systems. *Statistical Science* 8(3):219–283, 1993.

[Spirtes *et al.*, 1993] P. Spirtes, C. Glymour and R. Scheines. *Causation, Prediction, and Search (Lecture Notes in Statistics 81).* Springer, New York, NY, 1993.

[Spohn, 1980] W. Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic* 9:73–99, 1980.

[Thiele, 1997] W. Thiele. *Wasservögel und Strandvögel: Arten der Küsten und Feuchtgebiete*. BLV Naturführer, Munich, 1997.

[Verma and Pearl, 1990] T. S. Verma and J. Pearl. Causal networks: semantics and expressiveness. In: R. D. Shachter, T. S. Levitt L.N. Kanal, and J.F. Lemmer, eds. *Uncertainty in Artificial Intelligence 4*, pp. 69–76. North Holland, Amsterdam, 1990.

[Verma and Pearl, 1992] T. S. Verma and J. Pearl 1992. An algorithm for deciding if a set of observed independencies has a causal explanation. *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence (UAI'92, Stanford, CA)*, pp. 323–330. Morgan Kaufman, San Mateo, CA, 1992.

[Whittaker, 1990] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. J. Wiley & Sons, Chichester, 1990.

[Zhang and Poole, 1996] N. L. Zhang and D. Poole. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research* 5:301–328, 1996.