

# Probabilistic Argumentation Systems and Abduction \*

J. Kohlas (juerg.kohlas@unifr.ch), D. Berzati and R. Haenni  
*Department of Informatics DIUF*  
*University of Fribourg*  
*CH - 1700 Fribourg (Switzerland)*

**Abstract.** Probabilistic argumentation systems are based on assumption-based reasoning for obtaining arguments supporting hypotheses and on probability theory to compute probabilities of supports. Assumption-based reasoning is closely related to hypothetical reasoning or inference through theory formation. The latter approach has well known relations to abduction and default reasoning. In this paper assumption-based reasoning, as an alternative to theory formation aiming at a different goal, will be presented and its use for abduction and model-based diagnostics will be explained. Assumption-based reasoning is well suited for defining a probability structure on top of it. On the base of the relationships between assumption-based reasoning on the one hand and abduction on the other hand, the added value introduced by probability into model based diagnostics will be discussed. Furthermore, the concepts of complete and partial models are introduced with the goal to study the quality of inference procedures. In particular this will be used to compare abductive to possible explanations.

**Keywords:** Argumentation Systems, Assumption-Based Reasoning, Diagnostics, Belief Functions, Complete Models, Partial Models, Abduction.

## Table of Contents

1	Introduction	2
2	Complete and Partial Models	4
3	Possible Explanations	9
4	Abductive Explanations	12
5	Probabilistic Diagnoses	17
6	Conclusion	21

---

\* Research supported by grant No. 21-53500.98 of the Swiss National Foundation for Research.



## 1. Introduction

The idea of combining classical logic with probability theory leads to a more general theory of *probabilistic argumentation systems* (Anrig et al., 1999; Haenni, 1998; Kohlas and Monney, 1995). This theory is an alternative approach for non-monotonic reasoning under uncertainty. It allows to judge open questions (hypotheses) about the unknown or future world in the light of the given knowledge. From a qualitative point of view, the problem is to derive *arguments* in favor and against the hypothesis of interest. An argument can be seen as a set of possible assumptions that allows to deduce a hypothesis from a given knowledge base. Finally, a quantitative judgment of the situation is obtained by computing probabilities that the arguments are valid. The credibility of a hypothesis can then be measured by the total probability that it is supported or refuted by arguments. The resulting *degree of support* and *degree of possibility* correspond to (normalized) *belief* and *plausibility*, respectively, in the Dempster-Shafer theory of evidence (Kohlas and Monney, 1995; Shafer, 1976; Smets, 1998; Wilson, 1999). A quantitative judgment is often more useful and can help to decide whether a hypothesis can be accepted, rejected, or whether the available knowledge does not permit to decide.

Note that the concept of argumentation systems has many different meanings. We refer to (Chesñevar et al., 1998) for an overview. In this paper the notion of probabilistic argumentation systems is defined in Section 3.

In the past, some discussion about the appropriateness of probability and logic for common sense reasoning took place. In our view, there is not a competition between probability and logic. Both are needed for formalizing reasoning. This is no new view. Already George Boole used both logic and probability in his “Laws of Thought”. Our own combination of logic and probability extends this classical synthesis of the two domains. We use deduction or consequence finding to find arguments, in addition probability theory is used to measure the likelihood of arguments. This is similar in spirit to the approach of Neufeld and Poole (Neufeld and Poole, 1988) where probability theory is used to compare different possible theories. But our use of probability manifests the difference between assumption-based reasoning, as we understand it, and hypothetical reasoning and theory formation, as advocated for example by Poole (Poole, 1988a). As a consequence, probability theory leads in our framework in a very natural way to belief and plausibility functions in the sense of Dempster-Shafer theory. In this respect, probabilistic argumentation systems take up the work of (Laskey and Lehner, 1989; Provan, 1990; Pearl, 1988) and develop it further.

A fundamental property of the theory of probabilistic argumentation system is that additional knowledge may cause the judgment of the situation to change non-monotonically. Clearly, the property of *non-monotonicity* is required in any mathematical formalism for reasoning under uncertainty. It reflects a natural property of how a human's conviction or belief can change when new information is added. It may reinforce belief, but can also shed doubt on some previous belief. The theory of probabilistic argumentation systems shows that this kind of non-monotonicity can be achieved without leaving the field of classical logic. We illustrate this in this paper with respect to abduction and model-based diagnostics. In both cases the problem consist in finding explanations of the observed system behavior. The observations are compared with the expected system behavior and deviations from the expected behavior lead to assume possible abnormalities to explain the situation. The available knowledge about the underlying system is in general only partial. One of the goals of this paper is to study how the partiality of the knowledge of the system influences the results of inference or diagnosis.

For this purpose the notion of *complete models* is introduced. Such models are potential descriptions of the real system which generates the observations. Or, from an alternative point of view, they are models which allow a simulation of the process which generated the observation. In particular, there is a well defined state of the system, which induces observations. However, this state will be unknown and it is the purpose of the inference to find it out. It is assumed that the partial model used for inference is compatible with the complete model, which generated the observation. But the complete model is unknown in general; there will in fact be many complete models compatible with the partial model. It becomes then possible to judge the quality of the inference process with respect to the real state. This is similar to the theory of statistical inference, where properties of estimators and hypothesis tests like unbiasedness, variance, power of the test, etc. can be derived. This allows then to compare different estimators, tests, etc. and possibly to find optimal estimators and tests.

This study of the quality of the inference is almost completely missing for reasoning under uncertainty. This paper proposes a modest first step in this direction. In Section 2, the concepts of complete and partial models are introduced and the inference problem is posed. Possible explanations or diagnoses (Reiter, 1987; Darwiche, 1998; Darwiche, 2000) are defined and characterized in Section 3. It is shown that these explanations are at least complete in the sense that they contain the true state. In the same section, probabilistic argumentation systems are briefly introduced and it is shown how they relate to the inference

problem treated in this paper. The notion of possible explanations is then compared in Section 4 to the popular notion of abductive explanation as defined for example in (Kean, 1993; Poole, 1988a). It is shown that in general these latter notions are in a specific sense both incomplete and not sound for characterizing possible explanations. Finally, Section 5 explains the added value of obtaining both qualitative and quantitative supports for statements (hypotheses) about possible diagnoses with probabilistic argumentation systems. First of all, it is shown that the possible explanations constitute the sample space for the posterior probabilities, given the observations. So, for example in diagnostics, the posterior probability that specified components are faulty can be obtained. Maximum likelihood states are introduced and it is shown that they have a desirable property. This last section lies the base on which properties of state estimators (like the maximum likelihood state) and decision procedures can be studied. But this is a program which yet has to be executed.

## 2. Complete and Partial Models

When a diagnosis has to be made in a concrete situation, then it will be based on some knowledge about the system under consideration and on some observations regarding this system. Both the assumed knowledge as well as the observations will in general be incomplete. However, we assume that the elements used for the diagnosis will at least be consistent with the possibly unknown real system. Furthermore, we hope that the conclusion drawn from the partial knowledge and the incomplete observations will not be biased. But are these hopes really justified?

In order to examine this question, we introduce first the notion of a *complete model*, which is a potential description of real the system that generates the data used for the diagnosis. Then we define the notion of a *partial* model relative to a complete model. It represents an incomplete knowledge, which, however, is consistent with the underlying complete model. Inference or diagnosis will in general be based on partial models. We shall see that there are *several* complete models consistent with a given partial model. In practice, it will be unknown which one of the possible complete models corresponds to reality. But we may ask whether and to what extent a partial model yields reasonable results with respect to every consistent complete model. That is, we ask whether inference from partial models exhibits some kind of robustness.

Let's illustrate this by a very simple example. Figure 1 shows a single digital inverter. A *complete* model may state that (in propositional language)

$$ok \leftrightarrow (in \leftrightarrow \neg out).$$

This means that, if the inverter functions correctly (*ok*), then the output negates the input, whereas, if it does not function correctly, then the output equals the input. Why this can be considered to be a complete model will become clear below. A consistent partial model of this complete model would be

$$ok \rightarrow (in \leftrightarrow \neg out).$$

It tells us that the inverter negates the input if it functions properly. But it does not describe the system if the inverter is broken. The model now is only partial, but it is consistent with the original model, since it does not contradict it. A more precise definition of a partial model relative to a complete model is given below. As mentioned above, there are other complete models consistent with the partial model. For example, a model with a failure mode, in which *out* is always true, would be another one. Furthermore, a complete model with several possible failure states would be possible.

For the sake of simplicity, we limit the discussion, to propositional logic. Complete as well as partial models will therefore be described by propositional formulas. We consider two sets of propositional symbols  $A = \{a_1, \dots, a_m\}$  and  $P = \{p_1, \dots, p_m\}$ . The  $a_i$  are thought to indicate the states of the components, like the proposition *ok* in the example above. Typically,  $a_i$  means that the component  $i$  is properly functioning,  $\neg a_i$  that component  $i$  is broken. We shall assume that a probability  $\pi_i$  is associated with each component, describing its “reliability”. The propositions in  $P$  are the other symbols needed to formulate a system. Like *in* and *out* in the example above, some elements of the set  $P$  are thought to describe the “input” to the system and other elements of  $P$  describe the “output”.

$\mathcal{L}_{A \cup P}$  denotes the set of all well-formed propositional formulas over  $\mathcal{L}_{A \cup P}$ , where as usual,  $\wedge$  denotes conjunction,  $\vee$  disjunction,  $\neg$  negation,  $\rightarrow$  implication and  $\leftrightarrow$  equivalence. A *literal*  $l$  is a proposition or the negation of a proposition. A *term* is a conjunction of literals containing no repeated atoms.  $\mathcal{T}_Q$  denotes the set of all terms formed by literals of a set  $Q$  of propositions.  $\top$  denotes the “empty” conjunction. Often, we need terms which are of the form  $\ell_1 \wedge \ell_2 \wedge \dots \wedge \ell_k$ , where  $\ell_i = q$  or  $\ell_i = \neg q$  for  $i = 1, \dots, k$ ,  $q \in Q$  and  $k = |Q|$ , i.e. terms in  $\mathcal{T}_Q$  of maximal length  $|Q|$ . We call such terms *configurations* and denote their set by  $\mathcal{C}_Q$ .

A complete model is now a system, composed of components  $i = 1, 2, \dots, m$  whose state is described by proposition  $a_i$ . The *state* of the system is described by a configuration  $\ell_1 \wedge \ell_2 \wedge \dots \wedge \ell_m \in \mathcal{C}_A$ . Each  $\ell_i$  defines the state  $a_i$  or  $\neg a_i$  of the corresponding component. The behavior of the system is described by some formula  $\xi \in \mathcal{L}_{A \cup P}$ , which should describe how “output” is related to “input”.

Informally, the notion of a complete model is based on the following idea. If a system state is selected (for example by random sampling), then, for a given “input”, the “output” should be uniquely determined by  $\xi$ . That is, it should be possible to simulate the system behavior with a complete model. That is what the term “complete” refers to. A complete model should explain how observations (“inputs” and “outputs”), depending on the system state, are generated and related.

In order to define complete models more formally, we need to express the fact that the input can be chosen independently of the system state and the system description. The notion needed is defined as follows.

**DEFINITION 1.** *Let  $\xi \in \mathcal{L}_{A \cup P}$ . A subset  $Q \subseteq P$  of propositions is called logically independent of  $\xi$ , if for any state  $s \in \mathcal{C}_A$  for which  $s \wedge \xi$  is satisfiable, we have that  $\xi \wedge s \wedge \gamma$  is satisfiable for every configuration  $\gamma \in \mathcal{C}_Q$ . The logically independent set  $Q$  is called maximal, if no superset of  $Q$  is logically independent of  $\xi$ .*

Clearly, there may be several logically independent sets of  $\xi$ , even maximal ones.

*Example 1.* Consider the inverter of Figure 1. If the behavior of the inverter is modeled by  $\xi = ok \leftrightarrow (in \leftrightarrow out)$ , then the maximal logically independent sets are  $Q_1 = \{in\}$  and  $Q_2 = \{out\}$ . However,  $Q_3 = \{in, out\}$  is not logically independent of  $\xi$ .

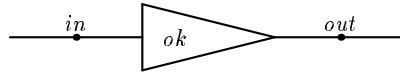


Figure 1. Inverter

From now on, if we speak of a logically independent set  $Q$ , we always assume that  $Q$  is maximal. Now, we are in a position to define the concept of a complete model.

**DEFINITION 2.** *A tuple  $(\hat{\xi}, A, P, \Pi, Q, R)$  is a complete model, if*

1.  $\hat{\xi} \in \mathcal{L}_{A \cup P}$ ;

2.  $\Pi$  is a set of stochastically independent probabilities  $\pi_i$  assigned to the propositions  $a_i$ ;
3.  $Q \subseteq P$  is logically independent of  $\hat{\xi}$ ;
4.  $R \subseteq P$  is disjoint of  $Q$ ;
5. for any system state  $\hat{s} \in \mathcal{C}_A$  such that  $\hat{s} \wedge \hat{\xi}$  is satisfiable, and for all configurations  $\hat{\gamma} \in \mathcal{C}_Q$ , there is a unique  $\hat{\tau} \in \mathcal{C}_R$  such that  $\hat{s} \wedge \hat{\xi} \wedge \hat{\gamma} \models \hat{\tau}$ .

The triple  $(\hat{s}, \hat{\gamma}, \hat{\tau})$  appearing in point (5) above is called a sample.

Here,  $\hat{\gamma}$  describes the “input” and  $\hat{\tau}$  the “output”. Point (5) of this definition requires then that for any admissible system state and any “input”  $\hat{\gamma}$  there is a uniquely determined “output”  $\hat{\tau}$ . This property allows to simulate the system behavior. Any simulation starts with a state  $\hat{s} \in \mathcal{C}_A$  (possibly obtained by random sampling using the probabilities in  $\Pi$ ) and an “input”  $\hat{\gamma} \in \mathcal{C}_Q$ . It produces then an “output”  $\hat{\tau} \in \mathcal{C}_R$ . That is why such a triple  $(\hat{s}, \hat{\gamma}, \hat{\tau})$  is called a sample.

*Example 2.* Assume a digital circuit containing three inverters with defined failure mode cabled in series as Figure 2 shows. Assume that

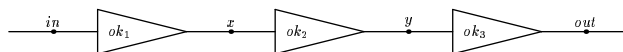


Figure 2. Three Not Gates

a faulty component passes only the received signal. The knowledge is then represented as  $\hat{\xi} = (ok_1 \leftrightarrow (in \leftrightarrow \neg x)) \wedge (ok_2 \leftrightarrow (x \leftrightarrow \neg y)) \wedge (ok_3 \leftrightarrow (y \leftrightarrow \neg out))$ .  $Q = \{in\}$  is a logically independent set of  $\hat{\xi}$ . Fix  $R = \{out\}$ . Clearly, for any state  $\hat{s}$  and input configuration  $\hat{\gamma}$ , the system description determines uniquely the output  $\hat{\tau}$ . For example,  $\hat{s} = ok_1 \wedge ok_2 \wedge ok_3$  and  $\hat{\gamma} = in$  imply  $\hat{\tau} = \neg out$ , whereas  $\hat{s} = \neg ok_1 \wedge ok_2 \wedge ok_3$  and  $\hat{\gamma} = in$  imply  $\hat{\tau} = out$ .

A complete model is a theoretical concept that models the background mechanism, which generates what we observe. Using Dupré’s (Dupré, 2000) terminology a complete model would be a *fully predictive model*. We assume that there is always a complete model that represents the real world. But we do not assume that the “real” complete model is known!

In fact, the actual knowledge about the real world is in general not complete. This means, that for purposes of inference, we may not know

the complete model that generated the available data. We have only partial information, which however is in a consistent way related to the complete model describing the real world.

**DEFINITION 3.** *Let  $(\hat{s}, \hat{\gamma}, \hat{\tau})$  be a sample of a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$ . A triple  $(\xi, \gamma, \tau)$  such that  $\xi \in \mathcal{L}_{A \cup P}$ ,  $\hat{\xi} \models \xi$ ,  $\hat{\gamma} \models \gamma$  and  $\hat{\tau} \models \tau$  is called an instance of an inference problem relative to the complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ . Furthermore,  $(\xi, A, P, \Pi)$  is called a partial model relative to the complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$ .*

So, in a partial model, we know the set of components  $A$  and the corresponding probabilities  $\Pi$ . The partial model is consistent with its underlying complete model, since  $\hat{\xi} \models \xi$ . But in general it is coarser. Also the complete “input”  $\hat{\gamma}$  and “output”  $\hat{\tau}$  are not necessarily known, only the partial “observations”  $\omega = \gamma \wedge \tau$ . The problem is to reconstruct the the actual state  $\hat{s}$  that generated  $\omega$  from the partial model  $\xi$  and the partial observation  $\omega$ . So, the complete model representing the “reality” behind the given data  $\xi$  and  $\omega$  is not known. We only claim that the inference problem is generated from such a complete model in the sense of Definition 3 above. This is illustrated in Figure 3. Note, once more,

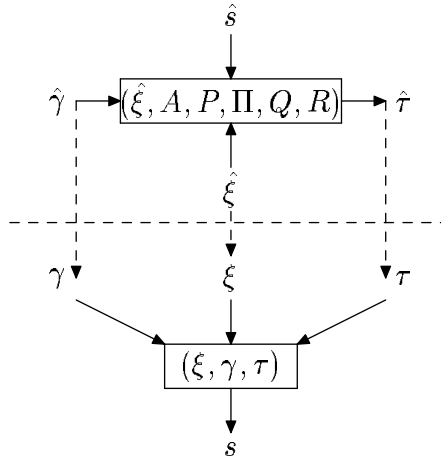


Figure 3. Inference Problem

that there can be many complete models and samples generating the same inference problem. The question is what can be found out about the state  $\hat{s}$ . We claim that this is, put into a propositional language, the problem of model-based diagnostics.

*Example 3.* Consider the same digital circuit as in Example 2, but with an undefined failure mode. Therefore, the only thing we know is the correct behavior of the components:  $\xi = (ok_1 \rightarrow (in \leftrightarrow \neg x)) \wedge (ok_2 \rightarrow (x \leftrightarrow \neg y)) \wedge (ok_3 \rightarrow (y \leftrightarrow \neg out))$ . Obviously,  $\hat{\xi} \models \xi$ , where  $\hat{\xi}$  is the system description of the complete model in Example 2. Thus,  $\xi$  forms a partial model relative to the complete model formed by  $\hat{\xi}$ . If we observe  $in \wedge out$ , then we have an instance of an inference problem  $(\xi, in, out)$ . The problem then is to find the state that generated this situation.

Note that in this example many different behavior under failure of the inverters can be imagined. This leads then really to different complete models. It indicates also, on an intuitive level, that it may be better to work with a good partial model than with the wrong complete model, which does not correspond to reality.

### 3. Possible Explanations

Consider an instance of an inference problem  $(\xi, \gamma, \tau)$  as defined in Section 2, Definition 3. The problem consists in finding the state  $\hat{s}$  that generated the available observation  $\omega = \gamma \wedge \tau$ . In general, even if  $\xi$  is a complete system description, there will be no unique solution to the inference problem. That is,  $\xi$ ,  $\gamma$  and  $\tau$  will not allow to reconstruct  $\hat{s}$  unambiguously. The notion of possible explanation delimits possible candidates for  $\hat{s}$ .

**DEFINITION 4.** *Let  $(\xi, \gamma, \tau)$  be an instance of an inference problem. A state  $s \in \mathcal{C}_A$  is a possible explanation of  $\omega = \gamma \wedge \tau$  with respect to  $\xi$ , if and only if  $\xi \wedge s \wedge \omega \not\models \perp$ .*

The idea behind this definition is simple to explain. If a state does not contradict what we observe, it should be accepted as a possible state that generated the observation. The relation  $\xi \wedge s \wedge \omega \not\models \perp$  is logically equivalent to  $\xi \wedge s \not\models \neg\omega$ .

*Example 4.* Consider the inverter of Example 1. Assume the system description is given by  $\xi = ok \leftrightarrow (in \leftrightarrow \neg out)$ , take  $Q = \{in\}$  as maximal logically independent set of  $\xi$ , and  $R = \{out\}$ . Assume the observation  $\omega = \neg in \wedge out$ . The  $s = ok$  is the only possible explanation. But suppose only the system description  $\xi = ok \rightarrow (in \leftrightarrow \neg out)$  is given. Then both  $ok$  and  $\neg ok$  are possible explanations.

The justification of the definition of possible explanations is given in the next theorem.

**THEOREM 1.** *Let  $(\xi, \gamma, \tau)$  be an instance of an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and generated by the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ . Then  $\hat{s}$  is a possible explanation of  $\gamma \wedge \tau$  with respect to  $\xi$ .*

*Proof.* Assume, there exists a state  $\hat{s} \in \mathcal{C}_A$  and a configuration  $\hat{\gamma} \in \mathcal{C}_Q$  such that  $\hat{s}$  and  $\hat{\gamma}$  generate  $\hat{\tau}$ , that is  $\hat{s} \wedge \hat{\xi} \wedge \hat{\gamma} \models \hat{\tau}$ .  $\hat{s} \wedge \hat{\xi} \wedge \hat{\gamma} \wedge \hat{\tau}$  is furthermore satisfiable. Since  $\hat{\xi} \models \xi$ ,  $\hat{\gamma} \models \gamma$  and  $\hat{\tau} \models \tau$ , we conclude that  $\hat{s} \wedge \xi \wedge \gamma \wedge \tau$  is satisfiable, hence,  $\hat{s} \wedge \xi \wedge \omega \not\models \perp$  (with  $\omega = \gamma \wedge \tau$ ). So  $\hat{s}$  is a possible explanation.

Therefore, by considering possible explanations, at least, we do not exclude the actual, unknown state that generated the observation. And this is true for any possible complete model that has possibly generated the instance of an inference problem. We say that the set of possible explanations is *complete*.

An other important requirement is that possible explanations do not contradict the partial system description  $\xi$ . This is excluded by Definition 4. However, the possible explanations may be inconsistent with an (unknown) complete model, which generated the instance of an inference problem. Therefore we say that the set of possible explanations is *weakly sound*. Weakly sound means that the possible explanations do not contradict the partial system description. Under what conditions possible explanations are sound with respect to a complete model will be worked out below.

At this point we want to link the inference problem to *assumption-based reasoning* (Haenni et al., 2000). We may consider the propositions of the set  $A$  as *assumptions* and  $\xi \in \mathcal{L}_{A \cup P}$  an assumption-based knowledge. A formula  $h \in \mathcal{L}_{A \cup P}$  may be considered as a hypothesis. We then want to know under what assumptions  $h$  can be deduced from  $\xi$ , or under what assumptions  $h$  can not be rejected based on the knowledge  $\xi$ . This kind of assumption-based reasoning is discussed in detail in (Haenni et al., 2000). There, states  $s \in \mathcal{C}_A$  are called *scenarios*. Here we will continue to call them states.

**DEFINITION 5.** *Let  $\xi \in \mathcal{L}_{A \cup P}$  and  $\Pi$  denote a set of probabilities  $\pi_i$  for all  $a_i \in A$ , then  $(\xi, A, P, \Pi)$  is called a probabilistic argumentation system.*

It is clear that a partial model as defined above can be considered as a probabilistic argumentation system. And this is not only a formal correspondence. There are important links between these two notions as will be explained below. The next definition distinguishes different kind of states.

**DEFINITION 6.** *A state  $s \in \mathcal{C}_A$  is called a*

1. inconsistent state relative to  $\xi$ , if  $s \wedge \xi$  is not satisfiable (we express this also as  $s \wedge \xi \models \perp$ );
2. consistent state relative to  $\xi$ , if  $s \wedge \xi$  is satisfiable;
3. quasi-supporting state for  $h$  relative to  $\xi$ , if  $s \wedge \xi \models h$ ;
4. supporting state for  $h$  relative to  $\xi$ , if  $s$  is consistent relative to  $\xi$  and if  $s \wedge \xi \models h$ ;
5. possible state for  $h$  relative to  $\xi$ , if  $s \wedge \xi \not\models \neg h$ .

Inconsistent states are to be considered as excluded by the knowledge  $\xi$ . Only consistent states remain possible in the light of  $\xi$ . Supporting states for  $h$  allow to deduce  $h$  from  $\xi$  and possible states for  $h$  do not allow to deduce  $\neg h$ , that is, do not allow to exclude  $h$ . Quasi-supporting states have a only technical and computational significance. We introduce corresponding sets

$$\begin{aligned}
 I_A(\xi) &= \{s \in \mathcal{C}_A : s \wedge \xi \models \perp\}, \\
 C_A(\xi) &= \{s \in \mathcal{C}_A : s \wedge \xi \not\models \perp\}, \\
 QS_A(h, \xi) &= \{s \in \mathcal{C}_A : s \wedge \xi \models h\}, \\
 SP_A(h, \xi) &= \{s \in \mathcal{C}_A : s \wedge \xi \models h\} \cap C_A(\xi), \\
 PS_A(h, \xi) &= \{s \in \mathcal{C}_A : s \wedge \xi \not\models \neg h\}.
 \end{aligned}$$

The sets of inconsistent and consistent states can also be expressed using quasi-supporting sets:  $I_A(\xi) = QS_A(\perp, \xi)$  and  $C_A(\xi) = \mathcal{C}_A - QS_A(\perp, \xi)$ . Furthermore, note that  $SP_A(h, \xi) \subseteq PS_A(h, \xi) \subseteq \mathcal{C}_A(\xi)$ .

The set of possible explanations of an inference problem  $(\xi, \gamma, \tau)$  may now be written as  $PS_A(\omega, \xi)$  with  $\omega = \gamma \wedge \tau$ . Note that  $SP_A(\omega, \xi) = \emptyset$ . Furthermore, observe that if the observation  $\omega$  is added to the knowledge  $\xi$  to obtain an enlarged knowledge  $\xi \wedge \omega$ , then  $\mathcal{C}_A(\xi \wedge \omega) = PS_A(\omega, \xi)$ . These sets of possible explanations may be very large and can in most cases not be listed explicitly. The important question of how to represent these large sets and of how to compute these representations is addressed in (Haenni et al., 2000).

#### 4. Abductive Explanations

In model-based diagnostics abductive explanations (or concepts derived from them) are often proposed as reasonable diagnoses. The usual definition of an *abductive explanation* (Poole, 1988a; Kean, 1993) is different from a possible explanation. A term  $\alpha \in \mathcal{T}_A$  of assumptions is

called an abductive explanation for an observation  $\omega = \gamma \rightarrow \tau$ , if the following two conditions are satisfied:

- (E1)  $\xi \wedge \alpha \models \omega$ ,
- (E2)  $\xi \wedge \alpha \not\models \perp$ .

Here, the observation is logically represented by a material implication  $\omega = \gamma \rightarrow \tau$  and not by a conjunction. Note that there are no abductive explanations for  $\omega = \gamma \wedge \tau$ . See (Poole, 1988b; Poole, 1989) for a detailed discussion.

*Example 5.* Consider the inverter of Example 1. Assume the inverter passes the received signal if it fails. Then, a possible complete system description is  $\hat{\xi} = ok \rightarrow (in \leftrightarrow \neg out) \wedge fail \rightarrow (in \leftrightarrow out) \wedge (ok \text{ xor } fail)$ , where xor denotes the exclusive or. If the observation is  $in \rightarrow out$ , then  $\alpha_1 = \neg ok$ ,  $\alpha_2 = fail$ , and  $\alpha_3 = \neg ok \wedge fail$  are the only three abductive explanations.

In order to discuss abductive explanations  $\alpha$  more generally, consider the states  $S_A(\alpha) := \{s \in \mathcal{C}_A : s \models \alpha\}$ , that is all states for which  $\alpha$  is true. The following theorem establishes the link between  $S_A(\alpha)$  and probabilistic argumentation systems.

**THEOREM 2.** *Let  $(\xi, P, A, \Pi)$  be a probabilistic argumentation system,  $\omega = \gamma \rightarrow \tau$ ,  $\gamma \in \mathcal{T}_Q$  and  $\tau \in \mathcal{T}_R$ , an observation,  $\alpha \in \mathcal{C}_A$  a term satisfying conditions (E1) and (E2) with respect to  $\xi$ , and  $S_A(\alpha)$  the set of corresponding states or scenarios, then*

- (1)  $S_A(\alpha) \subseteq QS_A(\omega, \xi)$ ,
- (2)  $S_A(\alpha) \cap SP_A(\omega, \xi) \neq \emptyset$ ,

*Proof.* Let  $\omega = \gamma \rightarrow \tau$  and  $\alpha \in \mathcal{C}_A$  a term satisfying (E1) and (E2), then

- (1) We know that  $\xi \wedge \alpha \models \omega$ , because of (E1). By definition  $s \models \alpha$  for all  $s \in S_A(\alpha)$ . From  $\xi \wedge \alpha \models \omega$  and  $s \models \alpha$  it follows that  $\xi \wedge s \models \omega$  for all states  $s \in S_A(\alpha)$ . Any states  $s \in S_A$  satisfying  $\xi \wedge s \models \omega$  is by definition a quasi-supporting state of  $\omega$ . Thus  $s \in QS_A(\omega, \xi)$  for all  $s \in S_A(\alpha)$  and so  $S_A(\alpha) \subseteq QS_A(\omega, \xi)$ .
- (2) Condition (E2) for  $\alpha$  guarantees that there exists a state  $\tilde{s} \in S_A(\alpha)$  such that  $\xi \wedge \tilde{s} \not\models \perp$ . We know by point (1) of Theorem 2 that  $\xi \wedge \tilde{s} \models \omega$ .  $\xi \wedge \tilde{s} \models \omega$  and  $\xi \wedge \tilde{s} \not\models \perp$  imply that  $\tilde{s}$  is a supporting state of  $\omega$ . Thus  $S_A(\alpha) \cap SP_A(\omega, \xi) \neq \emptyset$ .

The following example illustrates this theorem.

*Example 6.* Consider the inverter of Figure 1. Assume that the inverter is described in the same way as in Example 5. The states are  $s_1 = ok \wedge fail$ ,  $s_2 = ok \wedge \neg fail$ ,  $s_3 = \neg ok \wedge fail$  and  $s_4 = \neg ok \wedge \neg fail$ . Because the inverter can not be at the same time in ok mode and fail mode, the states  $s_1$  and  $s_4$  are inconsistent states. Then, from the results of Example 5 it follows that  $S_A(\alpha_1) = \{s_3, s_4\}$ ,  $S_A(\alpha_2) = \{s_1, s_3\}$ ,  $S_A(\alpha_3) = \{s_3\}$ ,  $QS_A(in \rightarrow out, \hat{\xi}) = \{s_1, s_3, s_4\}$ ,  $SP_A(in \rightarrow out, \hat{\xi}) = \{s_3\}$ . Obviously,  $S_A(\alpha_3) \subseteq S_A(\alpha_1) \subseteq QS_A(in \rightarrow out, \hat{\xi})$ ,  $S_A(\alpha_3) \subseteq S_A(\alpha_2) \subseteq QS_A(in \rightarrow out, \hat{\xi})$ , and  $S_A(\alpha_1) \cap SP_A(in \rightarrow out, \hat{\xi}) = S_A(\alpha_2) \cap SP_A(in \rightarrow out, \hat{\xi}) = S_A(\alpha_3) \cap SP_A(in \rightarrow out, \hat{\xi}) = \{s_3\}$ .

Another important relationship between abductive explanation and probabilistic argumentation system can be obtained by considering the set  $\Lambda(\omega, \xi)$  of all terms  $\alpha \in \mathcal{T}_A$  satisfying condition (E1) and (E2). If we consider now the states  $S_A(\Lambda(\omega, \xi)) := \bigcup \{S_A(\alpha) : \alpha \in \Lambda(\omega, \xi)\}$  we can deduce the following corollary.

**COROLLARY 1.** *Let  $(\xi, P, A, \Pi)$  be a probabilistic argumentation system,  $\omega = \gamma \rightarrow \tau$ ,  $\gamma \in \mathcal{T}_Q$  and  $\tau \in \mathcal{T}_R$ , an observation,  $\Lambda(\omega, \xi)$  the set of all terms  $\alpha \in \mathcal{C}_A$  satisfying conditions (E1) and (E2) with respect to  $\xi$  and  $\omega$ , and  $S_A(\Lambda(\omega, \xi))$  the corresponding states, then*

$$SP_A(\omega, \xi) \subseteq S_A(\Lambda(\omega, \xi)) \subseteq QS_A(\omega, \xi). \quad (1)$$

*Proof.* To prove  $SP_A(\omega, \xi) \subseteq S_A(\Lambda(\omega, \xi))$  consider a possible state  $\tilde{s} \in SP_A(\omega, \xi)$ . From this it follows immediately that  $\xi \wedge \tilde{s} \models \omega$  and  $\xi \wedge \tilde{s} \not\models \perp$ . Hence  $\tilde{s}$  is a (maximal) term satisfying conditions (E1) and (E2), so  $\tilde{s} \in S_A(\Lambda(\omega, \xi))$ .

The inclusion  $S_A(\Lambda(\omega, \xi)) \subseteq QS_A(\omega, \xi)$  follows from point (1) of Theorem 2, because  $S_A(\alpha) \subseteq QS_A(\omega, \xi)$  for any  $\alpha \in \Lambda(\omega, \xi)$  and thus  $S_A(\Lambda(\omega, \xi)) \subseteq QS_A(\omega, \xi)$ .

Thus, states supporting  $\omega$  relative to  $\xi$  are covered by abductive explanations. But there may be states, which belong to abductive explanations, but are not consistent relative to  $\xi$ . General abductive explanations  $\alpha$  are therefore not even weakly sound, in the sense that  $S_A(\alpha) \cap I_A(\xi)$  is not necessarily empty, as Example 6 shows.

In view of this conclusion, the requirement (E2) of abductive explanations, that is that an explanation need only be consistent with the available knowledge  $\xi$ , is not strong enough. One could think to augment the definition of abductive explanations to guarantee soundness. Change condition (E2) into

$$\xi \wedge \alpha' \not\models \perp, \quad \forall \alpha' : \alpha' \models \alpha. \quad (2)$$

This extension would resolve the problem of soundness of abductive explanations.

However, completeness should also be granted; we want to be sure not to miss the true state  $\hat{s}$  among the candidates. We show below that abductive explanations are in general not complete, except in special cases, namely when the inference is based on the complete model.

LEMMA 1. *Let  $(\hat{\xi}, \hat{\gamma}, \tau)$  be an instance of an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and generated by the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ , then*

$$\hat{s} \in SP_A(\hat{\gamma} \rightarrow \tau, \hat{\xi}). \quad (3)$$

*Proof.* We know that  $\hat{s} \in SP_A(\hat{\gamma} \rightarrow \hat{\tau}, \hat{\xi})$ , because  $\hat{\xi} \wedge \hat{s} \models \hat{\gamma} \rightarrow \hat{\tau}$  is equivalent to  $\hat{\xi} \wedge \hat{s} \wedge \hat{\gamma} \models \hat{\tau}$  and  $\hat{s}$  is a consistent state. From  $\hat{\tau} \models \tau$  we conclude that  $\hat{\xi} \wedge \hat{s} \wedge \hat{\gamma} \models \tau$ , and the proof is complete.

Thus, in the special case where the complete model *and* the complete “input”  $\hat{\gamma}$  is known, we can be sure that abductive explanations contain  $\hat{s}$ . We prove later, that this is the only case where abductive explanations assure completeness (see Theorem 3 and Theorem 4).

In the remainder of this section we are going to compare possible explanations with abductive explanations. To do so we have to state what an observation is. The following lemma shows that computing possible explanations for an observation  $\omega = \gamma \rightarrow \tau$  in the framework of probabilistic argumentation systems does not make much sense, since in this case the set of possible explanations corresponds to all consistent states relative to the partial model  $\xi$ .

LEMMA 2. *Let  $(\xi, \gamma, \tau)$  be an instance of an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and generated by the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ , then*

$$PS_A(\gamma \rightarrow \tau, \xi) = C_A(\xi). \quad (4)$$

*Proof.* Let  $\tilde{s} \in C_A(\xi)$  be any consistent state. We claim  $\xi \wedge \tilde{s} \wedge (\gamma \rightarrow \tau) \not\models \perp$ . This is logically equivalent to  $(\xi \wedge \tilde{s} \wedge \neg\gamma) \vee (\xi \wedge \tilde{s} \wedge \tau) \not\models \perp$ . Because each literal of  $\neg\gamma$  belongs to  $\mathcal{T}_Q$  and  $Q \subseteq P$  is a logically independent set of  $\hat{\xi}$  we have  $\hat{\xi} \wedge \tilde{s} \wedge \neg\gamma \not\models \perp$ . Together with  $\hat{\xi} \models \xi$  we get  $\xi \wedge \tilde{s} \wedge \neg\gamma \not\models \perp$ , and thus  $(\xi \wedge \tilde{s} \wedge \neg\gamma) \vee (\xi \wedge \tilde{s} \wedge \tau) \not\models \perp$ . So  $\xi \wedge \tilde{s} \wedge (\gamma \rightarrow \tau) \not\models \perp$  holds and  $\tilde{s} \in PS_A(\gamma \rightarrow \tau, \xi)$ . Since  $PS_A(\gamma \rightarrow \tau, \xi) \subseteq C_A(\xi)$ , we conclude  $PS_A(\gamma \rightarrow \tau, \xi) = C_A(\xi)$ .

Thus, if we speak of an observation  $\gamma$  and  $\tau$ , we use  $\gamma \wedge \tau$  when we compute possible explanations and  $\gamma \rightarrow \tau$  when we compute abductive explanations.

We know that abductive explanations for an observation  $\gamma \rightarrow \tau$  cover at least supporting states and at most quasi-supporting states (see Corollary 1). By definition, the supporting states are the consistent ones, because the quasi-supporting states can contain contradictory states. So the candidates for  $\hat{s}$  are the supporting states, if we compute abductive explanations for  $\gamma \rightarrow \tau$ . We can show that the possible explanations for  $\gamma \wedge \tau$  contain all consistent abductive explanations for  $\gamma \rightarrow \tau$ .

LEMMA 3. *Let  $(\xi, \gamma, \tau)$  be an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and generated by the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ , then*

$$SP_A(\gamma \rightarrow \tau, \xi) \subseteq PS_A(\gamma \wedge \tau, \xi). \quad (5)$$

*Proof.* Let  $s$  be any state in  $SP_A(\gamma \rightarrow \tau, \xi)$ . The following two conditions hold:  $\xi \wedge s \not\models \perp$  and  $\xi \wedge s \models \gamma \rightarrow \tau$ . We have the following chain of conclusions:

$$\begin{aligned} \xi \wedge s &\models \gamma \rightarrow \tau, \\ \xi \wedge s &\models \neg\gamma \vee \tau, \\ \xi \wedge s \wedge \gamma &\models \tau, \\ \xi \wedge s \wedge \gamma \wedge \tau &\not\models \perp. \end{aligned}$$

The last conclusion is valid since  $\xi \wedge s$  is satisfiable and  $Q$  logically independent of  $\xi$ . This is the definition of  $PS_A(\gamma \wedge \tau, \xi)$ . Therefore we conclude  $SP_A(\gamma \rightarrow \tau, \xi) \subseteq PS_A(\gamma \wedge \tau, \xi)$ .

Moreover, we can state that abductive explanations for  $\gamma \rightarrow \tau$  exclude in general candidates for  $\hat{s}$ . That is, abductive explanations are not complete. This is a very important result, which sheds serious doubts on the use of abductive explanations alone. Using only those diagnoses may mean to miss the true one. This is implied by the following theorem.

THEOREM 3. *Let  $(\xi, \gamma, \tau)$  be an instance of an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and generated by the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ , then*

$$S_A(\wedge(\gamma \rightarrow \omega, \xi)) \cap (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \rightarrow \tau, \xi)) = \emptyset. \quad (6)$$

*Proof.* Assume there exists an abductive explanation  $\alpha$  and a state  $\tilde{s} \in S_A(\alpha)$  such that  $\tilde{s} \in (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \rightarrow \tau, \xi))$ . By point (1) of Theorem 2  $\tilde{s} \in S_A(\alpha)$  implies that  $\xi \wedge \tilde{s} \models \gamma \rightarrow \tau$ , and this is equivalent to  $\xi \wedge \tilde{s} \wedge \gamma \models \tau$ .

On the other hand suppose  $\tilde{s} \in (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \rightarrow \tau, \xi))$ . This means that  $\xi \wedge \tilde{s} \wedge \gamma \wedge \tau \not\models \perp$  since  $\tilde{s} \in PS_A(\gamma \wedge \tau, \xi)$ . But  $\tilde{s} \notin SP_A(\gamma \rightarrow \tau, \xi)$ .

$\tau, \xi$ ) means that  $\xi \wedge \tilde{s} \not\models \gamma \rightarrow \tau$  since  $\xi \wedge \tilde{s}$  is satisfiable. This relation is equivalent to  $\xi \wedge \tilde{s} \wedge \gamma \not\models \tau$ . This contradicts  $\xi \wedge \hat{s} \wedge \gamma \models \tau$ . Thus, there is no state  $\tilde{s} \in S_A(\alpha)$  such that  $\tilde{s} \in (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \rightarrow \tau, \xi))$ .

The true state may be in the difference set  $PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \rightarrow \tau, \xi)$ . Therefore, the true state may not be covered by an abductive explanation. This is illustrated by the following example.

*Example 7.* Assume a simple digital circuit as Figure 4 shows.

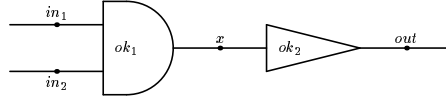


Figure 4. Simple Circuit

Let  $\xi = (ok_1 \rightarrow (in_1 \wedge in_2 \leftrightarrow x)) \wedge (ok_2 \leftrightarrow (x \leftrightarrow \neg out))$  be the system description of a partial model,  $\gamma = \neg in_1 \wedge \neg in_2$ , and  $\tau = \neg out$ . There are four states  $s_1 = ok_1 \wedge ok_2$ ,  $s_2 = ok_1 \wedge \neg ok_2$ ,  $s_3 = \neg ok_1 \wedge ok_2$ , and  $s_4 = \neg ok_1 \wedge \neg ok_2$ . The states  $s_2, s_3$ , and  $s_4$  are possible states for  $\neg in_1 \wedge \neg in_2 \wedge \neg out$ . There is only one abductive explanation, namely  $s_2$ , for  $\neg in_1 \wedge \neg in_2 \rightarrow \neg out$ . Clearly,  $S_A(\{s_2\}) = \{s_2\}$ .  $s_2$  is the supporting state for  $\neg in_1 \wedge \neg in_2 \rightarrow \neg out$ . Thus,  $SP_A(\neg in_1 \wedge \neg in_2 \rightarrow \neg out, \xi) \subset PS_A(\neg in_1 \wedge \neg in_2 \wedge \neg out, \xi)$  and  $S_A(\{s_2\}) \cap (PS_A(\neg in_1 \wedge \neg in_2 \wedge \neg out, \xi) - SP_A(\neg in_1 \wedge \neg in_2 \rightarrow \neg out, \xi)) = \emptyset$  hold.

We see that abductive explanations are in general not complete. If only a partial system description is given there is a chance that abductive explanations exclude the true state  $\hat{s}$  which generated what we observe. Nevertheless, Lemma 1 states that in the special case where the complete model *and* the configuration  $\hat{\gamma}$  is known abductive explanations contain  $\hat{s}$ . Not only that in this case abductive explanations are complete, but the consistent abductive explanations for an observation  $\hat{\gamma} \rightarrow \tau$  are exactly the possible explanations for the observation  $\hat{\gamma} \wedge \tau$  as the following theorem assures.

**THEOREM 4.** *Let  $(\hat{\xi}, \hat{\gamma}, \tau)$  be an instance of an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and generated by the sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ , then*

$$SP_A(\hat{\gamma} \rightarrow \tau, \hat{\xi}) = PS_A(\hat{\gamma} \wedge \tau, \hat{\xi}). \quad (7)$$

*Proof.* From Lemma 3 we know that  $SP_A(\hat{\gamma} \rightarrow \tau, \hat{\xi}) \subseteq PS_A(\hat{\gamma} \wedge \tau, \hat{\xi})$ . So we need only to prove the converse inclusion.

Let  $s \in PS_A(\hat{\gamma} \wedge \tau, \hat{\xi})$  which means that  $s \wedge \hat{\xi} \wedge \hat{\gamma} \wedge \tau \not\models \perp$ . Thus  $s \wedge \hat{\xi} \not\models \perp$ . Let  $\hat{\tau}$  be the output of the complete model associated with state  $s$  and input  $\hat{\gamma}$ ,  $s \wedge \hat{\xi} \wedge \hat{\gamma} \models \hat{\tau}$ .

We claim that  $\hat{\tau} \models \tau$ . For, if we assume that  $\hat{\tau} \not\models \tau$ , then  $\hat{\tau} \wedge \tau \models \perp$ , since  $\hat{\tau}$  is a maximal term in  $\mathcal{C}_R$ . But then  $s \wedge \hat{\xi} \wedge \hat{\gamma} \models \hat{\tau}$  implies  $s \wedge \hat{\xi} \wedge \hat{\gamma} \wedge \tau \models \perp$  contrary to the assumption that  $s \in PS_A(\hat{\gamma} \wedge \tau, \hat{\xi})$ .

So  $\hat{\tau} \models \tau$ , hence  $s \wedge \hat{\xi} \wedge \hat{\gamma} \models \tau$  which is equivalent to  $s \wedge \hat{\xi} \models \hat{\gamma} \rightarrow \tau$ . This, together with  $s \wedge \hat{\xi} \not\models \perp$ , shows that  $s \in SP_A(\hat{\gamma} \rightarrow \tau, \hat{\xi})$ .

To summarize, we can say that the notion of an abductive explanation is based on the idea, that an explanation, together with the knowledge base  $\xi$ , should imply the observed fact  $\omega = \gamma \rightarrow \tau$ . We think that this is too much asked. This condition makes only sense, if one can be sure to know the complete model of the given problem. On the other hand, it is usually required that an explanation is consistent with the available knowledge  $\xi$ . That, in our view, is not enough. We have seen, that this condition allows for inconsistencies.

So abductive explanations are in general neither complete nor sound, even if the observation is represented by an implication. That is why abduction in this sense is not fully appropriate for model-based diagnostics. In contrast, the notion of possible explanations in the framework of probabilistic argumentation systems fits perfectly all the requirements. And in particular it is exactly what is needed to introduce probabilities into the model. This will be discussed in the next section.

## 5. Probabilistic Diagnoses

The set of possible explanations of an instance of an inference problem may be very large. So, although we know that the “true” state  $\hat{s}$  is an element of this set, this may not be of much value. In this section, we show how the probability structure imposed on top of an argumentation system helps to get more information about  $\hat{s}$ . In fact, we may formulate different hypotheses about the unknown state, like that a given assumption holds or more complicated statements. And we may then compute the probabilities of such statements.

We must first define the probability space underlying the inference problems. The sample space is the set of all states  $\mathcal{C}_A$ . If  $\Pi$  is a set of probabilities  $\pi_i$  for each proposition  $a_i$  in  $A$ , then a state  $s = \ell_1 \wedge \ell_2 \wedge \dots \wedge \ell_m$  of  $\mathcal{C}_A$  obtains the probability

$$p(s) = \prod_{\ell_i = a_i} \pi_i \cdot \prod_{\ell_i = \neg a_i} (1 - \pi_i). \quad (8)$$

Here we assume that components fail independently from each other. This is not really an essential assumption for what follows. We may alternatively assume any probability distribution  $p(s)$  over  $\mathcal{C}_A$ . A subset  $S \subseteq \mathcal{C}_A$  obtains then the probability

$$P(S) = \sum_{s \in S} p(s). \quad (9)$$

This defines a probability space on the set of states. Note that this space is *independent* of the actual complete model we may want to consider (as long as the set  $A$  of components is fixed). If  $(\hat{\xi}, A, P, \Pi, Q, R)$  is a complete model,  $\hat{\gamma} \in \mathcal{C}_Q$  a fixed “input”, then we note, that the probability structure defined above induces a probability distribution on the “outputs”  $\hat{\tau} \in \mathcal{C}_R$ . In fact, for  $\hat{\xi}$  and  $\hat{\gamma}$  fixed, we have a mapping from  $\mathcal{C}_A$  into  $\mathcal{C}_R$  defined by  $s \rightarrow \hat{\tau}(s)$  such that  $s \wedge \hat{\xi} \wedge \hat{\gamma} \models \hat{\tau}(s)$ . Hence  $\hat{\tau}(s)$  is a random variable on this probability space.

Let now  $(\xi, \gamma, \tau)$  be an inference problem relative to a complete model  $(\hat{\xi}, A, P, \Pi, Q, R)$  and a sample  $(\hat{s}, \hat{\gamma}, \hat{\tau})$ . This is new data which restricts the possible states to a subset  $PS_A(\omega, \xi) = \mathcal{C}_A(\xi \wedge \omega) \subseteq \mathcal{C}_A$ . In terms of probability theory, this means that we should revise our prior probability defined by  $\Pi$  to conditional probabilities

$$p'(s) = p(s|\mathcal{C}_A(\xi \wedge \omega)) = \frac{p(s)}{p(\mathcal{C}_A(\xi \wedge \omega))} \text{ for } s \in \mathcal{C}_A(\xi \wedge \omega). \quad (10)$$

Thus,  $\mathcal{C}_A(\xi \wedge \omega)$  is the new sample space and the conditional probabilities  $p'(s)$  define a *posterior* probability space on  $\mathcal{C}_A(\xi \wedge \omega)$ . We shall use the convention that  $p'(s) = 0$ , if  $s \notin \mathcal{C}_A(\xi \wedge \omega)$ . We note again that this posterior probability depends only on the inference problem, but not on the underlying complete model. This is a lucky fact, since we may not know the actual complete model which generated the data.

Hypotheses about the unknown state  $\hat{s}$  which generated the inference problem may now be stated as subsets of the sample space  $\mathcal{C}_A(\xi \wedge \omega)$ . For example, the hypotheses that a fixed component  $i$  is working properly, that is that  $a_i$  holds, is represented by the set

$$\{s = \ell_1 \wedge \dots \wedge \ell_m \in \mathcal{C}_A(\xi \wedge \omega) : \ell_i = a_i\}.$$

The hypothesis that a fixed state  $s \in \mathcal{C}_A(\xi \wedge \omega)$  is the true state is simply defined by the single point set  $\{s\}$ . Hence, posterior probabilities of hypotheses  $H \subseteq \mathcal{C}_A(\xi \wedge \omega)$  can be computed as

$$p'(H) = \sum_{s \in H} p'(s). \quad (11)$$

Although this looks simple enough, we must emphasize that difficult computational problems may be associated with these posterior probabilities. This is due to the fact that sets like  $\mathcal{C}_A(\xi \wedge \omega)$  and  $H$  may

be very large. Already to compute conditional probabilities we must compute  $p(\mathcal{C}_A(\xi \wedge \omega))$ . These computational problems are addressed in (Kohlas et al., 1998) and will not be discussed here. We give simply an example to illustrate these probability computations in a very simple case.

*Example 8.* Consider Example 3, where the partial system description is given by  $\xi = (ok_1 \rightarrow (in \leftrightarrow \neg x)) \wedge (ok_2 \rightarrow (x \leftrightarrow \neg y)) \wedge (ok_3 \rightarrow (y \leftrightarrow \neg out))$ . Furthermore, assume that the (prior) probabilities are  $p(ok_1) = 0.97$ ,  $p(ok_2) = 0.93$ , and  $p(ok_3) = 0.95$ . We have the following eight states:

$$\begin{aligned} s_0 &= ok_1 \wedge ok_2 \wedge ok_3, & s_4 &= \neg ok_1 \wedge ok_2 \wedge ok_3, \\ s_1 &= ok_1 \wedge ok_2 \wedge \neg ok_3, & s_5 &= \neg ok_1 \wedge ok_2 \wedge \neg ok_3, \\ s_2 &= ok_1 \wedge \neg ok_2 \wedge ok_3, & s_6 &= \neg ok_1 \wedge \neg ok_2 \wedge ok_3, \\ s_3 &= ok_1 \wedge \neg ok_2 \wedge \neg ok_3, & s_7 &= \neg ok_1 \wedge \neg ok_2 \wedge \neg ok_3. \end{aligned}$$

The (prior) probability of the state  $s_0$  for example is  $p(s_0) = 0.97 \cdot 0.93 \cdot 0.95 = 0.857$ . If we observe  $in \wedge out$ , the state  $s_0$  is the one that is getting impossible, as we have seen. Conditioning on  $\mathcal{C}_A(\xi \wedge in \wedge out) = S_A - \{s_0\}$ , we obtain  $p(\mathcal{C}_A(\xi \wedge in \wedge out)) = 1 - p(s_0) = 0.143$ . So, the posterior probability for  $s_1$  for example becomes  $(0.97 \times 0.93 \times (1 - 0.95))/0.143 = 0.315$ .

These probabilities add value to the inference. They help to evaluate decision procedures to select diagnoses. For example, if  $\alpha \in \mathcal{T}_A$  is an abductive explanation for an observation  $\omega = \gamma \wedge \tau$ , then  $p'(S_A(\alpha))$  is the probability, that it hits the true state.  $p(S_A(\Lambda(\omega, \xi)))$  is the probability that we hit the true state with some abductive explanation. Since abductive explanations are not necessarily complete, this probability will in general be strictly smaller than one.

In the literature of model-based diagnostics several particular classes of diagnoses have been singled-out. We are now going to characterize probabilistically an important class introduced by Reiter (Reiter, 1987). If  $s$  is a possible explanation, we define  $s_-$  to be the set of negative assumptions in  $s$ . Furthermore, if  $s$  and  $s'$  are two states, we write  $s' \leq s$ , if  $s_- \subseteq s'_-$ . Reiter considers explanations which are minimal relative to this partial order in  $PS_A(\omega, \xi)$ . Especially in diagnostics, this makes much sense. It singles explanations out, for which the number of assumed faulty components (negative assumptions) is minimal.

LEMMA 4. *If for all assumptions  $a \in A$  we have that  $p(a) \geq p(\neg a) = 1 - p(a)$ , then for  $s, s' \in \mathcal{C}_A(\xi \wedge \omega)$  we have that  $s' \leq s$  implies  $p(s') \leq p(s)$ .*

*Proof.* Assume without loss of generality that  $s = \neg a_1 \wedge \dots \wedge \neg a_l \wedge a_{l+1} \wedge \dots \wedge a_m$ ,  $s' = \neg a_1 \wedge \dots \wedge \neg a_k \wedge a_{k+1} \wedge \dots \wedge a_m$ , and  $l \leq k$ . So  $p(s) = k' \cdot \prod_{i=1}^l (1 - p(a_i)) \cdot \prod_{i=l+1}^m p(a_i)$  and  $p(s') = k' \cdot \prod_{i=1}^k (1 - p(a_i)) \cdot \prod_{i=k+1}^m p(a_i)$ , where  $k' = p(C_A(\xi \wedge \omega))$ . We have to prove that  $p(s') \leq p(s)$ . We do this by direct calculation:

$$\begin{aligned}
p(s') &= k' \cdot \prod_{i=1}^k (1 - p(a_i)) \cdot \prod_{i=k+1}^m p(a_i) \\
&= k' \cdot \prod_{i=1}^l (1 - p(a_i)) \cdot \prod_{i=l+1}^k (1 - p(a_i)) \cdot \prod_{i=k+1}^m p(a_i) \\
&\leq k' \cdot \prod_{i=1}^l (1 - p(a_i)) \cdot \prod_{i=l+1}^k p(a_i) \cdot \prod_{i=k+1}^m p(a_i) \\
&= k' \cdot \prod_{i=1}^l (1 - p(a_i)) \cdot \prod_{i=l+1}^m p(a_i) = p(s).
\end{aligned}$$

The inequality holds, because  $1 - p(a_i) \leq p(a_i)$  for  $i = l + 1, \dots, k$ .

The assumption that  $p(a) \geq p(\neg a)$  is in diagnostics very reasonable, since the probability that a component functions correctly should (hopefully) be greater than the probability of a fault. From this lemma follows that Reiter explanations  $s$  have maximum probability among all possible states in their upsets  $s^\uparrow = ts' \in C_A(\xi \wedge \omega) : s' \geq s$ .

Of particular interest are *maximum likelihood* states. A state  $\tilde{s}$  is a maximum likelihood state, if

$$p(\tilde{s}|C_A(\xi \wedge \gamma \wedge \tau)) = \max_{s \in C_A(\xi \wedge \gamma \wedge \tau)} p(s|C_A(\xi \wedge \gamma \wedge \tau)) = k \cdot \max_{s \in C_A(\xi \wedge \gamma \wedge \tau)} p(s) \quad (12)$$

Note that there may be several maximum likelihood state. It follows immediately, that any maximum likelihood state is a Reiter explanation (but not the inverse). So, in order to search for the maximum likelihood states we need only to look through the Reiter states. Also it should be noted that the conditioning probability  $p(C_A(\xi \wedge \gamma \wedge \tau))$  needs not to be computed in order to find the maximum likelihood state. This is very important, because the computation of this probability can be difficult.

*Example 9.* Consider Example 8. The maximum likelihood state is  $s_2$ , with  $p(s_2) = 0.451$ . Indeed,  $s_2 = ok_1 \wedge \neg ok_2 \wedge ok_3$  is a Reiter explanation. It has a minimal number of faulty components.

## 6. Conclusion

In this paper we examine probabilistic argumentation systems as a way to combine logic and probability. Whereas in classical probability theory the logic part is limited to describe Boolean operations on events or Boolean connections of hypotheses, in argumentation systems logic plays a more important part. It is used to determine contradicting states and to delimit thus the remaining possible states. It is needed furthermore to find arguments in favor and against hypotheses. Probability on the other hand allows to compute the reliabilities of these arguments and to determine degrees of support and plausibilities of hypotheses. The classical probability measure on the space of assumptions is thereby extended to belief and plausibility functions (or in a more measure theoretic terminology to inner and outer measures) on the larger space of hypotheses.

The concept of probabilistic argumentation systems combines classical monotone logic and probability theory into a non-monotone inference formalism. This allows to consider various problems of common sense reasoning. In this paper we examined in particular abduction. Our approach allows to criticize the usual concept of abductive explanation. Furthermore, it adds value through the numerical evaluation of hypotheses by probabilities. Default logic and circumscription can be analyzed from the point of view of probabilistic argumentation systems in a similar way.

The notion of a complete model allows the study of the quality of inference and decision procedures. It could be studied by how far a state estimator (like the maximum likelihood state) can deviate from the true state, how likely faulty decisions (like replacing the wrong components) are, etc. This should allow to compare different inference and decision procedures and possibly to find optimal ones. May be the design of optimal experiments (for example which inputs to select) to get optimal inferences could also envisaged. These issues are yet to be studied.

## References

- Anrig, B., R. Bissig, R. Haenni, J. Kohlas, and N. Lehmann: 1999, 'Probabilistic Argumentation Systems: Introduction to Assumption-Based Modeling with ABEL'. Technical Report 99-1, University of Fribourg, Institute of Informatics, Theoretical Computer Science.
- Chesñevar, C., A. G. Maguitman, and R. P. Loui: 1998, 'Logical Models of Argument'. Draft for ACM Computing Surveys Logic of Arguments.
- Darwiche, A.: 1998, 'Model-based diagnosis using structured system descriptions'. *Artificial Intelligence Research* **8**, 165–222.

- Darwiche, A.: 2000, 'Model-Based Diagnosis under Real-World Constraints'. *The AI Magazine* **21**(2), 57–73.
- Dupré, D.: 2000, 'Abductive and Consistency-Based Diagnosis Revisited: a Modeling Perspective'. In: *Proc. of the 8th International Workshop on Non-Monotonic Reasoning, NMR'2000*.
- Haenni, R.: 1998, 'Modeling Uncertainty with Propositional Assumption-Based Systems'. In: S. Parson and A. Hunter (eds.): *Applications of Uncertainty Formalisms*, Lecture Notes in Artificial Intelligence 1455. Springer, pp. 446–470.
- Haenni, R., J. Kohlas, and N. Lehmann: 2000, 'Probabilistic argumentation systems'. In: J. Kohlas and S. Moral (eds.): *Defeasible Reasoning and Uncertainty Management Systems: Algorithms*. Oxford University Press.
- Kean, A.: 1993, 'The Approximation of Implicates and Explanations'. *International Journal of Approximate Reasoning* **9**, 97–128.
- Kohlas, J., B. Anrig, R. Haenni, and P. Monney: 1998, 'Model-Based Diagnostics and Probabilistic Assumption-Based Reasoning'. *Artificial Intelligence* **104**, 71–106.
- Kohlas, J. and P. Monney: 1995, *A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory of Evidence*, Vol. 425 of *Lecture Notes in Economics and Mathematical Systems*. Springer.
- Laskey, K. and P. Lehner: 1989, 'Assumptions, beliefs and probabilities'. *Artificial Intelligence* **41**, 65–77.
- Neufeld, E. and D. Poole: 1988, 'Combining logic and probability'. *Comp. Intelligence* **4**, 98–99.
- Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc.
- Poole, D.: 1988a, 'A Logical Framework for Default Reasoning'. *Artificial Intelligence* **36**, 27–47.
- Poole, D.: 1988b, 'Representing Knowledge for Logic-Based Diagnosis'. In: *Proc. International Conference of Fifth Generation Computer Systems*, pp. 1282–1290.
- Poole, D.: 1989, 'Normality and Faults in Logic-Based Diagnosis'. In: *Proc. Eleventh International Joint Conference on Artificial Intelligence*, pp. 1304–1310.
- Provan, G.: 1990, 'A Logic-based Analysis of Dempster-Shafer Theory'. *Int. J. Approximate Reasoning* **4**, 451–495.
- Reiter, R.: 1987, 'A Theory of Diagnosis from First Principles'. *Artificial Intelligence* **32**, 57–95.
- Shafer, G.: 1976, *The Mathematical Theory of Evidence*. Princeton University Press.
- Smets, P.: 1998, 'The Transferable Belief Model for Quantified Belief Representation'. In: D. Gabbay and P. Smets (eds.): *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, Vol. 1. Kluwer Academic Publishers, pp. 267–301.
- Wilson, N.: 1999, 'Algorithms for Dempster-Shafer Theory'. In: J. Kohlas and S. Moral (eds.): *Algorithms for Uncertainty and Defeasible Reasoning*. Kluwer Academic Publishers.

*Address for Offprints:* Department of Informatics DIUF  
 University of Fribourg  
 Rue de Faucigny 2  
 CH – 1700 Fribourg (Switzerland)