Linköping Electronic Articles in
Computer and Information Science
Vol. (): nr

Probabilistic Argumentation Systems and Abduction

Jürg Kohlas

Rolf Haenni

Dritan Berzati

Institute of Informatics
University of Fribourg
rue Faucigny 2
CH–1700 Fribourg, Switzerland
E–Mail: {juerg.kohlas, rolf.haenni, dritan.berzati}@unifr.ch

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis///

Published on , by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. (): nr .
http://www.ep.liu.se/ea/cis///. , .

This URL will also contain a link to the author's home page.

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.

Keywords: Argumentation Systems, Belief Functions,
Assumption-Based Reasoning, Abduction, Diagnostics.

1 INTRODUCTION

The idea of combining classical logic with probability theory leads to a more general theory of **probabilistic argumentation systems** [1, 5, 11]. 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 theory of evidence [11, 20, 21, 22]. 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.

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. Probability theory is used to measure the likelihood of arguments. This is similar in spirit to the approach of Neufeld and Poole [13] which introduce probability theory 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 [15]. 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 [12, 18, 14] 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. In section 2, the basic notions and concepts of probabilistic argumentation systems are introduced. The question arises, how probabilistic argumentation systems relate to other approaches of non-monotonic reasoning. In section 3, the case of **abduction** is discussed. Possible explanations or diagnoses [19, 2, 3] are defined and characterized. This notion is compared to the usual notion of abductive explanation as defined for example in [9, 15]. It is shown that these latter notions are in general in a specific sense both incomplete and not sound for characterizing possible explanations. In addition the added value of obtaining both qualitative and quantitative supports for statements (hypotheses) about possible diagnoses with probabilistic argumentation systems is explained. So, for example in diagnosis, the probability that specified components are faulty can be obtained.

2 PROBABILISTIC ARGUMENTATION SYSTEMS

Argumentation systems are obtained from propositional logic by considering two disjoint sets $P = \{p_1, \ldots, p_n\}$ and $A = \{a_1, \ldots, a_m\}$ of propositions. The elements of A are called **assumptions**. $\mathcal{L}_{A \cup P}$ denotes the corresponding propositional language. If ξ is a propositional sentence in $\mathcal{L}_{A \cup P}$, then a triple $\mathcal{AS} = (\xi, P, A)$ is called **propositional argumentation system**. ξ is called the **knowledge base** of $\mathcal{AS}$.

Sometimes, the knowledge base ξ is given as a conjunctive set $\Sigma = \{\xi_1, \ldots, \xi_r\}$ of sentences $\xi_i \in \mathcal{L}_{A \cup P}$ or, more specifically, clauses $\xi_i \in \mathcal{D}_{A \cup P}$. In such cases, it is always possible to use the corresponding conjunction $\xi = \xi_1 \wedge \cdots \wedge \xi_r$ instead. If $\xi \equiv \top$ (for example, if $\Sigma = \emptyset$), then ξ is called **vacuous** knowledge base. Similarly, ξ is called **contradictory**, if $\xi \equiv \bot$ (for example, if $\Sigma = \{\bot\}$).

2.1 Scenarios

The assumptions are essential for expressing uncertain information. They are used to represent uncertain events, unknown circumstances, or possible risks and outcomes. A conjunction $s = \ell_1 \wedge \cdots \wedge \ell_m$ of m literals $\ell_i = a_i$ or $\ell_i = \neg a_i$, $1 \leq i \leq m$, is called **scenario**. The set of all scenarios is denoted by $\mathcal{S}_A$. Scenarios represent possible **states** of the unknown or future world. This is the fundamental notion in this theory.

Some scenarios may become impossible with respect to the given knowledge base ξ. It is therefore necessary to distinguish two different types of scenarios. A scenario $s \in \mathcal{S}_A$ is called **inconsistent** (or **contradictory**) relative to ξ, iff $s \wedge \xi \models \bot$. Otherwise, s is called **consistent** relative to ξ.

Suppose that $s \in \mathcal{S}_A$ is an inconsistent scenario relative to ξ. This means that ξ becomes unsatisfiable when all the assumptions are set according to s. The set of all inconsistent scenarios is denoted by $I_A(\xi) = \{s \in \mathcal{S}_A : s \wedge \xi \models \bot\}$. Similarly, if s is supposed to be consistent relative to ξ, then ξ remains satisfiable when all the assumptions are set according to s. The set $C_A(\xi) = \{s \in \mathcal{S}_A : \ s \wedge \xi \not\models \bot\}$ denotes the collection of all consistent scenarios relative to ξ. Evidently, $I_A(\xi)$ and $C_A(\xi)$ are complementary set, that is $C_A(\xi) = \mathcal{S}_A - I_A(\xi)$.

Example 2.1.1. *Let $A = \{a_1, a_2\}$ and $P = \{p, q\}$ be two sets of propositions. If $\xi = (a_1 \rightarrow p) \wedge (a_2 \rightarrow q) \wedge (p \rightarrow \neg q)$ is a sentence in $\mathcal{L}_{A \cup P}$, then $I_A(\xi) = \{a_1 \wedge a_2\}$ is the set of inconsistent scenarios, and $C_A(\xi) = \{\neg a_1 \wedge \neg a_2, \neg a_1 \wedge a_2, a_1 \wedge \neg a_2\}$ is the set of consistent scenarios. The scenario $a_1 \wedge a_2$ is inconsistent because $a_1 \wedge a_2 \wedge \xi$ is unsatisfiable.*

The situation becomes more interesting when an additional propositional sentence $h \in \mathcal{L}_{A \cup P}$ called **hypothesis** is given. Hypotheses represent open questions or uncertain statements about some of the propositions in $A \cup P$. What can be inferred from ξ about the possible truth of h with respect to the given set of unknown assumptions? Possibly, if the assumptions are set according to some scenarios $s \in \mathcal{S}_A$, then h may be a logical consequence of ξ. In other words, h is supported by certain scenarios. More precisely, a scenario $s \in \mathcal{S}_A$ is called

(1) **quasi-supporting** scenario for h relative to ξ, iff $s \wedge \xi \models h$;

(2) **supporting** scenario for h relative to ξ, iff $s \wedge \xi \models h$ and $s \wedge \xi \not\models \bot$;

(3) **possible** scenario for h relative to ξ, iff $s \wedge \xi \not\models \neg h$.

The set $QS_A(h,\xi) = \{s \in \mathcal{S}_A : s \wedge \xi \models h\}$ denotes the collection of all quasi-supporting scenarios for h relative to ξ. Similarly, $SP_A(h,\xi) = \{s \in \mathcal{S}_A : s \wedge \xi \models h,\ s \wedge \xi \not\models \bot\}$ denotes the set of all supporting scenarios and $PS_A(h,\xi) = \{s \in \mathcal{S}_A : s \wedge \xi \not\models \neg h\}$ the set of all possible scenarios for h relative to ξ.

The difference between quasi-supporting and supporting scenarios is that quasi-supporting scenarios are allowed to be inconsistent. This is convenient for technical reasons. However, inconsistent scenarios are considered as impossible and hence excluded. Supporting scenarios are therefore more interesting.

Note that $SP_A(h,\xi) \subseteq QS_A(h,\xi)$ and $SP_A(h,\xi) \subseteq PS_A(h,\xi)$. Furthermore, $I_A(\xi) \subseteq QS_A(h,\xi)$, $SP_A(h,\xi) \subseteq C_A(\xi)$, and $PS_A(h,\xi) \subseteq C_A(\xi)$ for all hypotheses $h \in \mathcal{L}_{A \cup P}$.

Example 2.1.2. *Again, let $A = \{a_1, a_2\}$ and $P = \{p, q\}$. If $\xi = (a_1 \rightarrow p) \wedge (a_2 \rightarrow q) \wedge (p \rightarrow \neg q)$ is a sentence in $\mathcal{L}_{A \cup P}$, then $QS_A(p,\xi) = \{a_1 \wedge \neg a_2, a_1 \wedge a_2\}$ is the set of quasi-supporting scenarios for p, $SP_A(p,\xi) = \{a_1 \wedge \neg a_2\}$ the set of supporting scenarios for p, and $PS_A(p,\xi) = \{\neg a_1 \wedge \neg a_2, \neg a_1 \wedge a_2, a_1 \wedge \neg a_2\}$ the set of possible scenarios for p. Similarly, $QS_A(q,\xi) = \{\neg a_1 \wedge a_2, a_1 \wedge a_2\}$, $SP_A(q,\xi) = \{\neg a_1 \wedge a_2\}$, and $PS_A(q,\xi) = \{\neg a_1 \wedge \neg a_2, \neg a_1 \wedge a_2\}$ are the corresponding sets for q. Note that $a_1 \wedge a_2$ is inconsistent and therefore never a supporting or a possible scenario (see Example 2.1.1).*

The sets of inconsistent and consistent scenarios can be expressed in terms of quasi-supporting scenarios for $\bot$ by $I_A(\xi) = QS_A(\bot,\xi)$ and $C_A(\xi) = \mathcal{S}_A - QS_A(\bot,\xi)$, respectively. Similarly, the sets of supporting and possible scenarios for h can be determined via sets of quasi-supporting scenarios by $SP_A(h,\xi) = QS_A(h,\xi) - QS_A(\bot,\xi)$ and $PS_A(h,\xi) = \mathcal{S}_A - QS_A(\neg h,\xi)$, respectively. The problem of computing sets of inconsistent, consistent, supporting, and possible scenarios can therefore be solved by computing solely sets of quasi-supporting scenarios. Hence, the importance of quasi-supporting scenarios relies mainly on technical reasons, and also on the fact that $QS_A(h,\xi)$ can be determined more easily than $SP_A(h,\xi)$ [7].

An interesting situation to be considered is the case where the knowledge base ξ changes to $\xi' = \xi \wedge \tilde{\xi}$ by adding new information. Then, the number of inconsistent and quasi-supporting scenarios is monotonically increasing, whereas the number of consistent and possible scenarios is monotonically decreasing:

(1) $I_A(\xi') \supseteq I_A(\xi)$,

(2) $C_A(\xi') \subseteq C_A(\xi)$,

(3) $QS_A(h,\xi') \supseteq QS_A(h,\xi)$,

(4) $PS_A(h,\xi') \subseteq PS_A(h,\xi)$.

In contrast, nothing can be said about the number of supporting scenarios. If new information is added, then the set of supporting scenarios behaves **non-monotonically**, that is it may either grow or shrink, both cases are possible. The reason for this is that $SP_A(h,\xi)$ is a set difference of two monotonically growing sets $QS_A(h,\xi)$ and $QS_A(\bot,\xi)$.

The non-monotonicity of the set $SP_A(h,\xi)$ is an important property of argumentation systems. It reflects a natural property of how a human's conviction or belief can change when new information is given. Non-monotonicity is therefore a fundamental property for any mathematical formalism for reasoning under uncertainty. However, the concept of propositional argumentation systems shows that non-monotonicity can be achieved without leaving the field of classical monotone logic. As we will see, non-monotonicity is also observed when probabilities are assigned to the assumptions and conditional probabilities of $SP_A(h,\xi)$ given $C_A(\xi)$ are considered.

2.2 Arguments

Sets of scenarios $S \subseteq \mathcal{S}_A$ (such as $I_A(\xi)$, $C_A(\xi)$, $QS_A(h,\xi)$, $SP_A(h,\xi)$, and $PS_A(h,\xi)$) tend to grow exponentially with the size of A. An explicit representation as a list of elements $s \in S$ is therefore in general not feasible. Thus, alternative representations are needed. The idea is to represent the set S of scenarios by logical formulas. So S may always be described by the DNF (disjunctive normal form) $[S]_{\mathrm{dnf}} = \vee\{s \in S\}$. Every other logically equivalent DNF can then be used as an alternative representation of S. Obviously, short DNFs are of particular interest.

This way of representing sets of scenarios is known as **term representation** [7]. A **term** is a conjunction $\alpha = \ell_1 \wedge \cdots \wedge \ell_k$ of literals of non-repeating assumptions. $\mathcal{T}_A$ denotes the set of all terms, including the **empty term** $\top$. Note that every scenario is a term of maximal length with respect to A. Furthermore, every term is a sub-conjunction of one or several scenarios. If $T \subseteq \mathcal{T}_A$ is a set of terms, then a term $\alpha \in T$ is called **minimal** in T, if there is no other (shorter) term $\alpha' \in T$ such that $\alpha \models \alpha'$. μT denotes the corresponding set of minimal terms. Note that $[T]_{\mathrm{dnf}}$ and $[\mu T]_{\mathrm{dnf}}$ are logically equivalent.

A term $\alpha \in \mathcal{T}_A$ is called **inconsistent** (or **contradictory**) relative to ξ, iff $\alpha \wedge \xi \models \bot$. Furthermore, α is called **consistent** relative to ξ, iff $\alpha' \models \alpha$ implies $\alpha' \wedge \xi \not\models \bot$ for every $\alpha' \in \mathcal{T}_A$. There are in general terms $\alpha \in \mathcal{T}_A$ that are neither inconsistent nor consistent relative to ξ. The sets of all inconsistent and consistent terms are denoted by $I(\xi) = \{\alpha \in \mathcal{T}_A : \alpha \wedge \xi \models \bot\}$ and $C(\xi) = \{\alpha \in \mathcal{T}_A : \forall \alpha' \models \alpha,\ \alpha' \in \mathcal{T}_A,\ \alpha' \wedge \xi \not\models \bot\}$, respectively. $\mu I(\xi)$ and $\mu C(\xi)$ are the corresponding sets of minimal terms. Note that $[\mu I(\xi)]_{\mathrm{dnf}} \equiv [I_A(\xi)]_{\mathrm{dnf}}$ and $[\mu C(\xi)]_{\mathrm{dnf}} \equiv [C_A(\xi)]_{\mathrm{dnf}}$.

If a hypothesis $h \in \mathcal{L}_{A\cup P}$ is given, then a term $\alpha \in \mathcal{T}_A$ is called **quasi-supporting argument** for h relative to ξ, iff $\alpha \wedge \xi \models h$. Furthermore, α is called **supporting argument** for h relative to ξ, iff (1) α is a quasi-supporting argument for h and (2) α is consistent relative to ξ. Finally, α is a **possible argument** for h relative to ξ, iff $\alpha' \models \alpha$ implies $\alpha' \wedge \xi \not\models \neg h$ for every $\alpha' \in \mathcal{T}_A$. The sets of all quasi-supporting, supporting, and possible arguments are denoted by $QS(h,\xi) = \{\alpha \in \mathcal{T}_A : \alpha \wedge \xi \models h\}$, $SP(h,\xi) = \{\alpha \in \mathcal{T}_A : \alpha \wedge \xi \models h,\ \forall \alpha' \models \alpha,\ \alpha' \in \mathcal{T}_A,\ \alpha' \wedge \xi \not\models \bot\}$, and $PS(h,\xi) = \{\alpha \in \mathcal{T}_A : \forall \alpha' \models \alpha,\ \alpha' \in \mathcal{T}_A,\ \alpha' \wedge \xi \not\models \neg h\}$, respectively. $\mu QS(h,\xi)$, $\mu SP(h,\xi)$, and $\mu PS(h,\xi)$ are the corresponding sets of minimal terms. Note that $[\mu QS(\xi)]_{\mathrm{dnf}} \equiv [QS_A(\xi)]_{\mathrm{dnf}}$, $[\mu SP(\xi)]_{\mathrm{dnf}} \equiv [SP_A(\xi)]_{\mathrm{dnf}}$ and $[\mu PS(\xi)]_{\mathrm{dnf}} \equiv [PS_A(\xi)]_{\mathrm{dnf}}$.

2.3 Degrees of Support and Possibility

So far, the problem of judging hypotheses has only been considered from a qualitative point of view. A more sizable judgment of the hypothesis can be obtained, if a probability measure on the set $N_A = \{0,1\}^m$ of interpretations of the language $\mathcal{L}_A$ is defined by $p(\mathbf{x})$ for all $\mathbf{x} \in N_A$. $\Pi = \{p(\mathbf{x}) : \mathbf{x} \in N_A\}$ denotes the set of all such probabilities. One way of obtaining such a probability measure is to assign independent probabilities π_i to the assumption $a_i \in A$. In such a case, the probability of an interpretation $\mathbf{x} = (x_1, \ldots, x_m)$ is given by

$$p(\mathbf{x}) \;=\; \prod_{i=1}^{m} {\pi_i}^{x_i} \cdot (1-\pi_i)^{(1-x_i)}. \tag{1}$$

The quadruple $\mathcal{PAS} = (\xi, P, A, \Pi)$ is called **probabilistic argumentation system**.

Note that every scenario $s \in \mathcal{S}_A$ has a corresponding interpretation in $\mathbf{x}_s \in N_A$ and vice versa. Therefore, it is possible to write $p(s) = p(\mathbf{x}_s)$. If $S \subseteq \mathcal{S}_A$ is an arbitrary set of scenarios, then the probability of S is simply the sum of the probabilities of its elements:

$$p(S) \;=\; \sum_{s \in S} p(s). \tag{2}$$

If $h \in \mathcal{L}_{A \cup P}$ is a hypothesis to be judged, then

$$dqs(h,\xi) \;=\; p(QS_A(h,\xi)) \tag{3}$$

is called **degree of quasi-support** of h relative to ξ. This measure corresponds to **unnormalized belief** in the Dempster-Shafer theory of evidence [20, 21].

$p(s)$ is to be considered as a prior probability. The knowledge ξ defines an event $C_A(\xi)$. In view of this new information, the probability has to be conditioned as usual in probability theory. This leads to the new probability measure p' given by

$$p'(\mathbf{x}) \;=\; p(\mathbf{x}|C_A(\xi)) \;=\; \begin{cases} \frac{p(\mathbf{x})}{p(C_A(\xi))}, & \text{if } \mathbf{x} \in C_A(\xi), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The prior probabilities $p(\mathbf{s})$ of consistent scenarios are multiplied by a **normalization factor** $k = p(C_A(\xi))^{-1}$. Note that $p(C_A(\xi))$ is the same as $1 - dqs(\perp, \xi)$. The new probability measure p' defines **posterior** probabilities for the scenarios given the knowledge base ξ. If $h \in \mathcal{L}_{A \cup P}$ is a hypothesis, then

$$\begin{aligned} dsp(h,\xi) \;&=\; p'(SP_A(h,\xi)) \;=\; \sum_{\mathbf{s} \in SP_A(h,\xi)} p'(\mathbf{s}) \\ &=\; \frac{1}{p(C_A(\xi))} \cdot \sum_{\mathbf{s} \in SP_A(h,\xi)} p(\mathbf{s}) \;=\; \frac{p(SP_A(h,\xi))}{p(C_A(\xi))} \\ &=\; \frac{p(QS_A(h,\xi)) - p(QS_A(\perp,\xi))}{1 - p(QS_A(\perp,\xi))} \\ &=\; \frac{dqs(h,\xi) - dqs(\perp,\xi)}{1 - dqs(\perp,\xi)} \end{aligned} \tag{5}$$

is called **degree of support** of h relative to ξ. It corresponds to **normalized belief** in the Dempster-Shafer theory of evidence. Note that $dsp(h,\xi) = p(SP_A(h,\xi)|C_A(\xi))$. Degree of support can therefore be considered as the conditional probability of $SP_A(h,\xi)$ given $C_A(\xi)$. If the knowledge base ξ is contradictory, then $dsp(h,\perp)$ is undefined. An important property of $dsp(h,\xi)$ is that it behaves non-monotonically when new knowledge is added.

A second posterior measure for hypotheses is obtained by considering the corresponding conditional probability on the set of possible scenarios $PS_A(h,\xi)$. Therefore,

$$\begin{aligned} dps(h,\xi) &= p'(PS_A(h,\xi)) = \sum_{\mathbf{s}\in PS_A(h,\xi)} p'(\mathbf{s}) \\ &= \frac{1}{p(C_A(\xi))} \cdot \sum_{\mathbf{s}\in PS_A(h,\xi)} p(\mathbf{s}) \;=\; \frac{p(PS_A(h,\xi))}{p(C_A(\xi))} \\ &= \frac{1-p(QS_A(\neg h,\xi))}{1-p(QS_A(\perp,\xi))} \;=\; \frac{1-dqs(\neg h,\xi)}{1-dqs(\perp,\xi)} \\ &= 1-dsp(\neg h,\xi) \end{aligned} \tag{6}$$

is called **degree of possibility**. Again, $dps(h,\perp)$ is undefined. The corresponding notion in the context of the Dempster-Shafer theory is **plausibility**. Also here, a non-monotonic behavior is observed when new knowledge is added.

An important property follows from the fact that $SP_A(h,\xi)$ is always a subset of $PS_A(h,\xi)$:

$$dsp(h,\xi) \;\leq\; dps(h,\xi). \tag{7}$$

All this indicates that the framework of probabilistic argumentation systems constructed on propositional logic is a alternative formalism for Shafer's original evidence theory [20].

The formalism of probabilistic argumentation systems can be extended to the more general approach of set constraint logic [8], where the assumptions can be multi-valued.

3 ABDUCTION

An important feature of probabilistic argumentation systems is the fact that it supports several kinds of non-monotonic inference. One form of non-monotonic reasoning is logic-based abduction, which is applicable when the available knowledge can be represented as a logical theory. Abduction is, in general terms, inference from a given knowledge ξ to a possible explanation for an observed or reported fact τ.

We place this problem in the framework of assumption-based reasoning and probabilistic argumentation systems. In this view, the assumptions describe, as we have seen, uncertain events or statements. Scenarios describe states of the real world. Together with a given knowledge about the world, scenarios determine possible outcomes or observations. We are first going to formalize this mechanism. The problem is then to infer from the knowledge and the observations the unknown state of the world, i.e. the scenario which generated the observation.

3.1 Complete and Partial Models

We start by describing a mechanism which generates the data available for the subsequent inference. There is a logical **systems description** $\hat{\xi} \in \mathcal{L}_{A \cup P}$. Depending on the actual scenario $\hat{s}$, this system generates some observations. We may think of these observations as "input" and "output" values of the system. Often, the "input" values can be freely selected and only then the "output" values are determined. That is the intuition behind the concepts that will be introduced now.

Remember that a term is a conjunction of non-repeating literals over a given alphabet. If $\mathcal{T}_A$ denotes the set of all possible terms over A, then the set of scenarios $\mathcal{S}_A \subseteq \mathcal{T}_A$ denotes the set of terms of maximal length. To describe "input" and "output" of a system, we need to consider two disjoint (sub-) alphabets $Q \subseteq P$ and $R \subseteq P - Q$. The sets of corresponding terms are then $\mathcal{T}_Q$ and $\mathcal{T}_R$, respectively. We denote the set of terms with maximal length over Q and R by $\mathcal{C}_Q$ and $\mathcal{C}_R$, respectively, and call them **configurations**. If we fix a consistent scenario $\hat{s} \in C_A(\xi)$ and a configuration $\hat{\gamma}$ ("input") , then these two elements, together with the system description $\hat{\xi}$ determine unambiguously the configuration in R ("output"). To formalize this idea, we need the following definitions:

Definition 3.1.1. *Let ξ be a formula from $\mathcal{L}_{A \cup P}$. A subset $Q \subseteq P$ of propositions is called* **logically independent** *of ξ, if for any consistent scenario $s \in C_A(\xi)$ and any configuration $\gamma \in \mathcal{C}_Q$ the formula $\xi \wedge s \wedge \gamma$ is consistent. The logically independent set Q is called* **maximal**, *if there is no other logically independent subset $\tilde{Q} \subseteq P$ with $Q \subseteq \tilde{Q}$.*

Clearly, the set Q is not unique, even the maximal set is not unique.

Example 3.1.1. *Consider an inverter as Figure 1 shows. 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 ξ.*

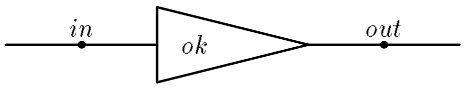

Figure 1: Inverter

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. It is thought to be a system description, which describes how observations are generated.

Definition 3.1.2. *Let $(\hat{\xi}, A, P, \Pi)$ be a probabilistic argumentation system and $Q \subseteq P$ a logically independent set of $\hat{\xi}$, $R \subseteq P$ disjoint of Q. Then $(\hat{\xi}, A, P, \Pi, Q, R)$ is called a* **complete model**, *if for any consistent scenario $s \in C_A(\hat{\xi})$ and all configurations $\gamma \in \mathcal{C}_Q$ there is a unique $\tau \in \mathcal{C}_R$ such that $\hat{\xi} \wedge s \wedge \gamma \models \tau$.*

We call the triple (s, γ, τ) a **sample** *of the complete model.*

Associated with a complete model is a set of possible **observations** $O(\hat{\xi}, Q, R) := \{\omega = \gamma \wedge \tau \in \mathcal{T}_{Q \cup R} : \exists s \in C_A(\hat{\xi}) \text{ such that } \hat{\xi} \wedge s \wedge \gamma \models \tau\}$.

A complete model is a theoretical concept that models the background mechanism, which generates what we observe. In Dupré's [4] terminology a complete model is a *fully predictive model*. We assume that there exists always a complete model which represents the real world.

Example 3.1.2. *Assume a digital circuit containing three inverters with defined failure mode cabled in series as Figure 2 shows. Assume that 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* $\hat{\xi}$. *If we fix* $R = \{out\}$, *then the observations are* $O(\hat{\xi}, Q, R) = \{in \wedge out, in \wedge \neg out, \neg in \wedge out, \neg in \wedge \neg out\}$. *Clearly, each possible observation* $\omega \in O(\hat{\xi}, Q, R)$ *is a consequence of the system description, a certain scenario* s, *and a configuration* $\gamma \in \mathcal{C}_Q$. *For example,* $s_1 = ok_1 \wedge ok_2 \wedge ok_3$ *and* $\gamma = in$ *imply* $\neg out$ *whereas* $s_2 = \neg ok_1 \wedge ok_2 \wedge ok_3$ *and* $\gamma = in$ *imply* out.

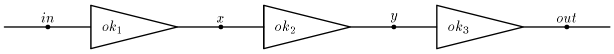

Figure 2: Three Not Gates

The actual knowledge about the real world however is in general not complete. This means, that for purposes of inference, we may not know the complete model which generated the available data. We have only partial information, which however is somehow related to the complete model describing the real world. What we mean by this is defined as follows:

Definition 3.1.3. *Let* $(\hat{\xi}, A, P, \Pi, Q, R)$ *be a complete model and* $(\hat{s}, \hat{\gamma}, \hat{\tau})$ *a corresponding sample. A triple* (ξ, γ, τ) *such that* $\xi \in \mathcal{L}_{A \cup P}$ *and* $\hat{\xi} \models \xi$, $\gamma \in \mathcal{T}_Q$ *and* $\hat{\gamma} \models \gamma$, $\tau \in \mathcal{T}_R$ *and* $\hat{\tau} \models \tau$ *is called an* **instance** *of an* **inference problem** *relative to the complete model* $(\hat{\xi}, A, P, \Pi, Q, R)$.

We call ξ a **partial system description**. $\omega = \gamma \wedge \tau$ is the observation available for the inference. The problem consists in inferring from the available knowledge (ξ, γ, τ) the actual scenario $\hat{s}$, which generated the observation ω. The whole idea is illustrated by the picture in Figure 3. There is the complete model, which, according to a scenario $\hat{s}$ and a selected "input" configuration $\hat{\gamma}$, generates an "output" configuration $\hat{\tau}$. These elements in turn imply the actual observation $\omega = \gamma \wedge \tau$, that is available for the inference. The inference has to be done using a partial system description ξ and it results somehow in a scenario s, which is a possible explanation of the given data (see subsection 3.2).

We illustrate this with the inverter system.

Example 3.1.3. *Consider the same digital circuit as in Example 3.1.2, but with undefined failure mode. Therefore, the only thing we know is the behavior of the components if they would work correctly:* $\xi = (ok_1 \rightarrow (in \leftrightarrow \neg x)) \wedge (ok_2 \rightarrow (x \leftrightarrow \neg y)) \wedge (ok_3 \rightarrow (y \leftrightarrow \neg out))$. *Obviously,* $\hat{\xi} \models \xi$, *where* $\hat{\xi}$ *is the system description of the complete model in Example 3.1.2. Thus, this is a partial system description. If we observe* $in \wedge out$, *then we have an instance of an inference problem* (ξ, in, out) *and we would like to find the scenario that generated this situation.*

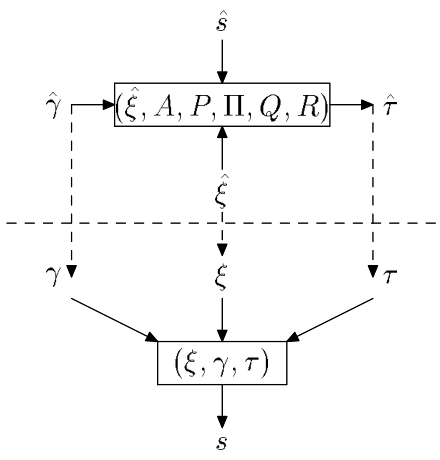

Figure 3: Inference Problem

The definition of a partial system description ξ is introduced because in general the actual knowledge is often not complete. So we don't ask the complete model to be known. We only claim that there is such a complete model and that any actual instance of an inference problem (ξ, γ, τ) is generated from a complete model. Note also that there can be several complete models associated to an instance of an inference problem. Example 3.1.3 shows this clearly since we may complete the system description by defining several, different failure modes.

3.2 Possible Explanation

Consider an instance of an inference problem (ξ, γ, τ) as defined in subsection 3.1. The problem consists in finding the scenario $\hat{s}$ that generated the available observation $\omega = \gamma \wedge \tau$. In general, even if ξ is a complete system description, there will be no unique solution to the inference problem. That is, ξ, γ and τ will not allow to reconstruct $\hat{s}$ unambiguously.

The notion of possible explanation delimits possible candidates for $\hat{s}$.

Definition 3.2.1. *Let (ξ, γ, τ) be an instance of an inference problem. A scenario $s \in \mathcal{S}_A$ is a (possible)* ***explanation*** *of $\omega = \gamma \wedge \tau$ iff*

$$\xi \wedge s \wedge \omega \not\models \perp.$$

The idea behind this definition is simple to explain. If a scenario does not contradict what we observe, it should be accepted as a possible scenario that generated the observation. The relation $\xi \wedge s \wedge \omega \not\models \perp$ is logically equivalent to $\xi \wedge s \not\models \neg\omega$. But this is exactly the definition of the possible scenarios for ω (see subsection 2.1). Hence, the set of possible explanations for the observation ω is exactly the set of possible scenarios $PS_A(\omega, \xi)$ for ω.

Example 3.2.1. *Consider the inverter of Example 3.1.1. Assume the system description is given by $\xi = ok \leftrightarrow (in \leftrightarrow out)$, take $Q = \{in\}$ as maximal logically independent set of ξ, and $R = \{out\}$. The possible observations are $O(\xi, Q, R) = \{in \wedge out, in \wedge \neg out, \neg in \wedge out, \neg in \wedge \neg out\}$. Assume in to be false and out to be true, then $PS_A(\neg in \wedge out, \xi) = \{ok\}$. If out would be false, then $PS_A(\neg in \wedge \neg out, \xi) = \{\neg ok\}$.*

The definition of a possible explanation uses scenarios. As we mentioned in subsection 2.2, the concept of term representation for sets of scenarios is often useful. Using the definition of possible arguments for ω, we get $PS(\omega, \xi) = \{\alpha \in \mathcal{T}_A : \forall \alpha' \models \alpha, \alpha \in \mathcal{T}_A, \alpha' \wedge \xi \not\models \neg(\omega)$. Note that the assumptions not appearing in $\alpha \in \mathcal{T}_A$ can have any value. This means, they are irrelevant to ω. Remember that term representation and scenario representation are equivalent in the framework of probabilistic argumentation systems.

The justification for our definition is given in the following theorem.

Theorem 3.2.1. *Let (ξ, γ, τ) 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 ξ.*

Proof. Assume there exists a scenario $\hat{s} \in \mathcal{C}_A(\hat{\xi})$ and a configuration $\hat{\gamma} \in \mathcal{C}_Q$ such that that $\hat{s}$ and $\hat{\gamma}$ generate $\hat{\tau} \in \mathcal{C}_R$ with respect to $\hat{\xi}$ but $\hat{s} \notin PS_A(\omega, \xi)$, where $\omega = \gamma \wedge \tau$. If $\hat{s} \notin PS_A(\omega, \xi)$ then $\hat{s} \in QS_A(\neg\omega, \xi)$ and so $\xi \wedge \hat{s} \models \neg\omega$. Last statement is logically equivalent to $\xi \wedge \hat{s} \wedge \omega \models \perp$. But we know that $\hat{\xi} \wedge \hat{s} \wedge \hat{\omega} \not\models \perp$, where $\hat{\omega} = \hat{\gamma} \wedge \hat{\tau}$. This is clearly a contradiction, because $\hat{\gamma} \models \gamma$ and $\hat{\tau} \models \tau$ and hence $\hat{\omega} \models \omega$. We have proven ad absurdum that all generating scenario for any observation ω with respect to the complete model is included in $PS_A(\omega, \xi)$ and so $\hat{s} \in PS_A(\omega, \xi)$. □

So, at least, by considering possible explanations, we do not exclude the actual, unknown scenario, which generated the observation. And this is true for any possible complete model, which generated possibly 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 ξ. This is excluded by definition of the possible explanations. 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 the complete model has still to be worked out.

We resume that looking for possible explanations is complete and weakly sound.

3.3 Abductive Explanation

The usual definition of an **abductive explanation** [15, 9] is different: 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) $\xi \wedge \alpha \models \omega$, and

(E2) $\xi \wedge \alpha \not\models \perp$.

The observation is logically represented by a material implication, and not by a conjunction. See [16, 17] for detailed discussion, and note that the implication is crucial for the further discussion.

Example 3.3.1. *Consider the inverter of Example 3.1.1. Assume the inverter passes the received signal if it fails. Then the system description is* *given by* $\hat{\xi} = ok \rightarrow (in \leftrightarrow out) \wedge fail \rightarrow (in \leftrightarrow out) \wedge (ok \text{ xor } fail)$. *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 α more generally, consider the models $S_A(\alpha) := \{s \in \mathcal{S}_A : s \models \alpha\}$, that is, all scenarios, which make α true. The following theorem establishes the link between $S_A(\alpha)$ and probabilistic argumentation systems.

Theorem 3.3.1. *Let* (ξ, P, A, Π) *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* ξ, *and* $S_A(\alpha)$ *the set of corresponding scenarios, then*

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

Proof. Let $\omega = \gamma \rightarrow \tau$ be a fact 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 $s \in S_A(\alpha)$ holds. From $\xi \wedge \alpha \models \omega$ and $s \models \alpha$ it follows that $\xi \wedge s \models \omega$ for all scenario $s \in S_A(\alpha)$. Any scenario $s \in S_A$ satisfying $\xi \wedge s \models \omega$ is by definition a quasi-supporting scenario of ω. 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 α guarantees that there exists a scenario $\tilde{s} \in S_A(\alpha)$ such that $\xi \wedge \tilde{s} \not\models \bot$. We know by point (1) of theorem 3.3.1 that $\xi \wedge \tilde{s} \models \omega$. $\xi \wedge \tilde{s} \models \omega$ and $\xi \wedge \tilde{s} \not\models \bot$ imply that $\tilde{s}$ is a supporting scenario of ω. Thus $S_A(\alpha) \cap SP_A(\omega, \xi) \neq \emptyset$.

□

This theorem says essentially that computing abductive explanations α corresponds to computing the quasi-support of the hypothesis $\omega = \gamma \rightarrow \tau$ in the framework of probabilistic argumentation systems (point (1)). Moreover, the abductive explanations α include scenarios which are possible candidates for $\hat{s}$ (point (2) of the theorem).

Example 3.3.2. *Consider the inverter of Figure 1. Assume that the inverter passes the received signal, if it fails, as in Example 3.3.1. The scenarios 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 cannot be at the same time in ok mode and fail mode, the scenarios* s_1 *and* s_3 *are inconsistent scenarios. Assume the observation is given by* $in \rightarrow out$, *then* $\alpha_1 = \neg ok$, $\alpha_2 = fail$ *and* $\alpha_3 = \neg ok \wedge fail$ *are abductive explanations for* $in \rightarrow out$ *with respect to* $\hat{\xi}$. *We can compute* $S_A(\alpha_1) = \{s_3, s_4\}$, $S_A(\alpha_2) = \{s_1, s_3\}$, $S_A(\alpha_3) = \{s_3\}$, $QS_A(in \rightarrow out, \hat{\xi}) = \{s_1, s_3, s_4\}$, $SP_A(in \rightarrow out, \hat{\xi}) = \{s_3\}$. *Obviously,* $S_A(\alpha_2) \subseteq S_A(\alpha_1) \subseteq QS_A(in \rightarrow out, \hat{\xi})$ *and* $S_A(\alpha_1) \cap SP_A(in \rightarrow out, \hat{\xi}) = S_A(\alpha_2) \cap SP_A(in \rightarrow out, \hat{\xi}) = \{s_3\}$.

Note that Theorem 3.3.1 considered the models of any abductive explanation α and the relationships to the corresponding scenarios. Another important relationship between abductive explanation and probabilistic argumentation system can be drawn by considering the set Λ_α of all terms $\alpha \in \mathcal{T}_A$ satisfying condition (E1) and (E2). If we consider now the models $S_A(\Lambda_\alpha) := \bigcup\{S_A(\alpha) : \alpha \in \Lambda_\alpha\}$ we can deduce following corollary.

Corollary 3.3.1. *Let* (ξ, P, A, Π) *be a probabilistic argumentation system,* $\omega = \gamma \to \tau$*,* $\gamma \in \mathcal{T}_Q$ *and* $\tau \in \mathcal{T}_R$*, an observation,* Λ_α *the set of all terms* $\alpha \in \mathcal{C}_A$ *satisfying conditions* (E1) *and* (E2) *with respect to* ξ*, and* $S_A(\Lambda_\alpha)$ *the corresponding models, then*

$$SP_A(\omega, \xi) \subseteq S_A(\Lambda_\alpha) \subseteq QS_A(\omega, \xi).$$

Proof. To prove $SP_A(\omega, \xi) \subseteq S_A(\Lambda_\alpha)$ assume there exists a scenario $\tilde{s} \in SP_A(\omega, \xi)$ such that $\tilde{s} \notin S_A(\Lambda_\alpha)$. From $\tilde{s} \in SP_A(\omega, \xi)$ follows immediately that $\xi \wedge \tilde{s} \models \omega$ and $\xi \wedge \tilde{s} \not\models \perp$. On the other hand Λ_α contains all terms $\alpha \in \mathcal{C}_A$, such that $\xi \wedge \alpha \models \omega$ and $\xi \wedge \alpha \not\models \perp$. $\tilde{s}$ is a (maximal) term satisfying conditions (E1) and (E2), so $\tilde{s} \in S_A(\Lambda_\alpha)$. This contradicts the assumption $\tilde{s} \notin S_A(\Lambda_\alpha)$. So all scenarios $s \in SP_A(\omega, \xi)$ are contained in $S_A(\Lambda_\alpha)$.

The inclusion $S_A(\Lambda_\alpha) \subseteq QS_A(\omega, \xi)$ follows from point (1) of Theorem 3.3.1, because $S_A(\alpha) \subseteq QS_A(\omega, \xi)$ for any $\alpha \in \Lambda_\alpha$ and thus $S_A(\Lambda_\alpha) \subseteq QS_A(\omega, \xi)$. □

So the only consistent scenarios in $S_A(\Lambda_\alpha)$ correspond to the supporting scenarios in the framework of probabilistic argumentation systems. But in general abductive explanations are not sound, that is the cut $S_A(\alpha) \cap I_A(\xi)$ is not empty as Example 3.3.3 shows.

Example 3.3.3. *Consider the system description of Example 3.3.1* $\hat{\xi} = ok \to (in \leftrightarrow out) \wedge fail \to (in \leftrightarrow out) \wedge (ok \text{ xor } fail)$*, and the observation* $in \to out$*. There are four scenarios:* $s_1 = ok \wedge fail$*,* $s_2 = ok \wedge \neg fail$*,* $s_3 = \neg ok \wedge fail$*, and* $s_4 = \neg ok \wedge \neg fail$*. The term* $\alpha = \neg ok$ *is an abductive explanation of* $in \to out$*. But* $s_4 \in S_A(\alpha)$ *and* s_4 *is an inconsistent scenario, that is* $s_4 \in I_A(\xi)$*.*

In our view, the requirement (E2) of abductive explanations, that is an explanation is only consistent with the available knowledge ξ, is not enough. One could think to augment the definition of abductive explanations to guarantee soundness. Change condition (E2) into

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

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

However, inference should also guarantee completeness. We will show that abductive explanations are in general not complete, except in special cases. Those special cases are the ones we know to built the complete model, that generated the actual observation. Then we can state following lemma:

Lemma 3.3.1. *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*

$$\hat{s} \in SP_A(\hat{\gamma} \to \tau, \hat{\xi}).$$

Proof. We know that $\hat{s} \in SP_A(\hat{\gamma} \to \hat{\tau}, \hat{\xi})$, because $\hat{\xi} \wedge \hat{s} \models \hat{\gamma} \to \hat{\tau}$ is logically equivalent to $\hat{\xi} \wedge \hat{s} \wedge \hat{\gamma} \models \hat{\tau}$ and $\hat{s}$ is a consistent scenario. 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 configuration $\hat{\gamma}$ is known one 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 3.3.2 and Theorem 3.3.3).

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 $\omega = \gamma \to \tau$ in the framework of probabilistic argumentation systems does not make much sense.

Lemma 3.3.2. *Let* (ξ, γ, τ) *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*

$$PS_A(\gamma \to \tau, \xi) \;=\; C_A(\xi).$$

Proof. Let $\tilde{s} \in C_A(\xi)$ be any consistent scenario. To show $\xi \wedge \tilde{s} \wedge (\gamma \to \tau) \not\models \perp$. This is logically equivalent to $(\xi \wedge \tilde{s} \wedge \neg\gamma) \vee (\xi \wedge \tilde{s} \wedge \tau) \not\models \perp$. Because of $\neg\gamma \in \mathcal{T}_Q$ and $Q \subseteq P$ 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 $\xi \wedge \tilde{s} \wedge (\gamma \to \tau) \not\models \perp$ holds. We conclude $PS_A(\gamma \to \tau, \xi) = C_A(\xi)$. ☐

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

We know that abductive explanations for an observation $\gamma \to \tau$ cover at least supporting scenarios and at most quasi-supporting scenarios (see Corollary 3.3.1). By definition, the supporting scenarios are the consistent ones, because the quasi-supporting scenarios can contain contradictory scenarios. So the candidates for $\hat{s}$ are supporting scenarios, if we compute abductive explanations for $\gamma \to \tau$. We can prove that possible explanations for $\gamma \wedge \tau$ contain all consistent abductive explanations for $\gamma \to \tau$.

Lemma 3.3.3. *Let* (ξ, γ, τ) *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*

$$SP_A(\gamma \to \tau, \xi) \subseteq PS_A(\gamma \wedge \tau, \xi).$$

Proof. Let s be any scenario in $SP_A(\gamma \to \tau, \xi)$. The following two conditions hold: $\xi \wedge s \not\models \perp$ and $\xi \wedge s \models \gamma \to \tau$. By direct manipulation we get:

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

And this is the definition of $PS_A(\gamma \wedge \tau, \xi)$. By the choice of s we conclude $SP_A(\gamma \to \tau, \xi) \subseteq PS_A(\gamma \wedge \tau, \xi)$. ☐

Moreover, we can state that abductive explanations for $\gamma \to \tau$ exclude in general candidates for $\hat{s}$.

Theorem 3.3.2. *Let* (ξ, γ, τ) *be an instance of an inference problem relative to a complete model* $(\hat{\xi}, A, P, \Pi, Q, R)$ *and generated by the sample* $(\hat{s}, \hat{\gamma}, \hat{\tau})$*,* $\alpha \in \mathcal{C}_A$ *such that (1)* $\xi \wedge \alpha \models \gamma \to \tau$ *and (2)* $\xi \wedge \alpha \not\models \perp$*, and* $S_A(\alpha)$ *the set of corresponding scenarios, then*

$$S_A(\alpha) \cap (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \to \tau, \xi)) = \emptyset.$$

Proof. Assume there exists a scenario $\tilde{s} \in S_A(\alpha)$ such that $\tilde{s} \in (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \to \tau, \xi))$. By point (1) of Theorem 3.3.1 we know that $\xi \wedge \tilde{s} \models \gamma \to \tau$, and this is equivalent to $\xi \wedge \tilde{s} \wedge \gamma \models \tau$.

The relation $\tilde{s} \in PS_A(\gamma \wedge \tau, \xi)$ implies that $\xi \wedge \tilde{s} \wedge \gamma \wedge \tau \not\models \bot$. On the other hand $\xi \wedge \tilde{s} \not\models \gamma \to \tau$, because the set of supporting scenarios is excluded. This is equivalent to $\xi \wedge \tilde{s} \wedge \gamma \not\models \tau$.

We know that $\xi \wedge s \wedge \gamma \not\models \bot$ for any consistent scenario $s \in C_A(\xi)$ (see Definition 3.1.2: logical independence). A scenario $\tilde{s}$ which $\xi \wedge \tilde{s} \wedge \gamma \models \tau$ and $\xi \wedge \tilde{s} \wedge \gamma \not\models \tau$ is a contradictory scenario. The scenarios in $(PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \to \tau, \xi))$ are not contradictory. Thus there cannot be a scenario $\tilde{s} \in S_A(\alpha)$ such $\tilde{s} \in (PS_A(\gamma \wedge \tau, \xi) - SP_A(\gamma \to \tau, \xi))$. □

Example 3.3.4. *Assume a simple digital circuit as Figure 4 shows.*

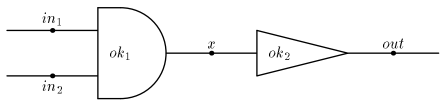

Figure 4: Simple Circuit

Let $\xi = (ok_1 \to (in_1 \wedge in_2 \leftrightarrow x)) \wedge (ok_2 \leftrightarrow (x \leftrightarrow \neg out))$ *be the system description,* $\gamma = \neg in_1 \wedge \neg in_2$*, and* $\tau = \neg out$*. There are four scenarios* $s_1 = ok_1 \wedge ok_2$*,* $s_2 = ok_1 \wedge \neg ok_2$*,* $s_3 = \neg ok_1 \wedge ok_2$*, and* $s_4 = \neg ok_1 \wedge \neg ok_2$*. The scenarios* s_2, s_3*, and* s_4 *are possible scenarios for* $\neg in_1 \wedge \neg in_2 \wedge \neg out$*. There is only one abductive explanation, namely* s_2*, for* $\neg in_1 \wedge \neg in_2 \to \neg out$*. Clearly,* $S_A(s_2) = s_2$ *holds.* s_2 *is the supporting scenario for* $\neg in_1 \wedge \neg in_2 \to \neg out$*. Thus* $SP_A(\neg in_1 \wedge \neg in_2 \to \neg out, \xi) \subseteq PS_A(\neg in_1 \wedge \neg in_2 \wedge \neg out, \xi)$ *and* $S_A(s_2) \cap (PS_A(\neg in_1 \wedge \neg in_2 \wedge \neg out, \xi) - SP_A(\neg in_1 \wedge \neg in_2 \to \neg out, \xi)) = \emptyset$ *hold.*

We see that abductive explanations are in general not complete. If only a partial system description is given there is a high chance that abductive explanations exclude the true scenario $\hat{s}$ which generated what we observe. Nevertheless, Lemma 3.3.1 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 $\hat{\gamma} \to \tau$ are exactly the possible explanations for the observation $\hat{\gamma} \wedge \tau$ as following theorem assures.

Theorem 3.3.3. *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*

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

Proof. We know that the logically independent set $Q \subseteq P$ is maximal and $\hat{\gamma} \in \mathcal{C}_Q$ a configuration. This means that for all consistent scenario $\tilde{s} \in C_A(\hat{\xi})$ there exists a configuration $\tilde{\tau} \in \mathcal{C}_R$ such that $\hat{\xi} \wedge \tilde{s} \wedge \hat{\gamma} \models \tilde{\tau}$. Let $C_A(\hat{\xi}) = \{s_1, \ldots, s_k\}$ be the set of consistent scenarios and $O(\hat{\gamma}) = \{\tau_1, \ldots, \tau_k\}$ the corresponding configurations, such that $\hat{\xi} \wedge s_i \wedge \hat{\gamma} \models \tau_i$, $i = 1, \ldots, k$. Let $C_A^+(\hat{\xi}) := \{s_i \in C_A(\hat{\xi}) : \tau_i \models \tau\}$ and $C_A^-(\hat{\xi}) := \{s_i \in C_A(\hat{\xi}) : \tau_i \not\models \tau\}$. Because $\tilde{\tau} \in \mathcal{C}_R$ is maximal, it follows immediately that $\tilde{\tau} \wedge \tau \models \bot$ for all $\tau \in \mathcal{T}_R$ such that $\tilde{\tau} \not\models \tau$. It is evident that $C_A^+(\hat{\xi}) \subseteq SP_A(\hat{\gamma} \to \tau, \hat{\xi})$, and $C_A^-(\hat{\xi}) \cap SP_A(\hat{\gamma} \to \tau, \hat{\xi}) = \emptyset$ hold. From

Lemma 3.3.3 we know that $SP_A(\hat{\gamma} \to \tau, \hat{\xi}) \subseteq PS_A(\hat{\gamma} \wedge \tau, \hat{\xi})$, and thus $C_A^+(\hat{\xi}) \subseteq PS_A(\hat{\gamma} \wedge \tau, \hat{\xi})$. Rests $C_A^-(\hat{\xi}) \cap PS_A(\hat{\gamma} \wedge \tau, \hat{\xi}) = \emptyset$ to prove. But for any scenario $\tilde{s} \in C_A^-(\hat{\xi})$ it holds that $\tilde{\tau} \wedge \tau \models \perp$ and hence $\hat{\xi} \wedge \tilde{s} \wedge \hat{\gamma} \wedge \tau \models \perp$, because $\hat{\xi} \wedge \tilde{s} \wedge \hat{\gamma} \models \tilde{\tau}$. From $\hat{\xi} \wedge \tilde{s} \wedge \hat{\gamma} \wedge \tau \models \perp$ we conclude $\hat{\xi} \wedge \tilde{s} \models \neg(\hat{\gamma} \wedge \tau)$, and thus $\tilde{s} \in PS_A(\neg(\hat{\gamma} \wedge \tau), \hat{\xi})$. So $C_A^-(\hat{\xi}) \cap PS_A(\hat{\gamma} \wedge \tau, \hat{\xi}) = \emptyset$ holds and the proof is complete. □

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

We ask only that an explanation does not contradict the knowledge ξ, including the observed fact $\omega = \gamma \wedge \tau$. In other words, we require that an explanation is consistent, not only with the available information ξ, but also with the observation $\omega = \gamma \wedge \tau$. With this simple condition we are sure to include the true scenario $\hat{s}$.

So abductive explanations are in general neither complete nor sound, even if the observation is represented by a implication. That is why abduction in this sense is not appropriate for model based diagnostics. In contrast, the notion of possible explanations in the framework of probabilistic argumentation systems fits perfectly all the requirements.

3.4 Probabilistic Explanations

The set of possible explanations of an instance of an inference problem may by very large. So, although we know that the "true" scenario $\hat{s}$ 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 scenario, like that a given assumption holds or more complicated statements. And we may then compute the probabilities of such statements.

We assume that $(\hat{\xi}, P, A, \Pi, Q, R)$ is a complete model and (ξ, P, A, Π) where $\hat{\xi} \models \xi$ is a probabilistic argumentation system with the same assumptions and the same probabilities as the complete model. That is, if we have an instance (ξ, γ, τ) of an inference problem relative to the complete model $(\hat{\xi}, P, A, \Pi, Q, R)$ and generated by a sample $(\hat{s}, \hat{\gamma}, \hat{\tau})$, then we assume that we know the (prior) probabilities $p(a_i)$ of the assumptions.

Now, the observation $\omega = \gamma \wedge \tau$ is additional information to ξ. Therefore, we consider the probabilistic argumentation system $(\xi \wedge \omega, P, A, \Pi)$. As we have seen in subsection 2.3, this defines a conditional probability p' on the set $C_A(\xi \wedge \omega)$. Note that $C_A(\xi \wedge \omega) = PS_A(\omega, \xi)$. So p' is the posterior probability on the possible explanations. This is underlined by the following theorem proved in [10, 6]:

Theorem 3.4.1. *If $h_A \in \mathcal{L}_A$ and $\xi \in \mathcal{L}_{A \cup P}$, then*

(1) $QS_A(h_A, \xi') = S_A(h_A) \cup I_A(\xi')$,

(2) $SP_A(h_A, \xi') = PS_A(h_A, \xi') = S_A(h_A) \cap C_A(\xi')$,

So, according to this theorem, if $h \in \mathcal{L}_A$ is any hypothesis about possible scenarios, then we have

$$\begin{aligned} dsp(h, \xi \wedge \omega) &= dpl(h, \xi \wedge \omega) \\ &= p(S_A(h)|C_A(\xi \wedge \omega)). \end{aligned} \tag{8}$$

From a computational point of view, this means essentially that the probability $p(C_A(\xi \wedge \omega)) = 1 - p(I_A(\xi \wedge \omega))$ must be computed, since

$$p(h|C_A(\xi \wedge \omega) = k \cdot p(S_A(h) \cap C_A(\xi \wedge \omega)), \tag{9}$$

where

$$k^{-1} = p(C_A(\xi \wedge \omega)). \tag{10}$$

In particular, the posterior probabilities of possible scenarios are

$$p(s|C_A(\xi \wedge \omega) = k \cdot p(s). \tag{11}$$

Example 3.4.1. *Consider Example 3.1.3, 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 scenarios:*

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

The (prior) probability of the scenario 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 scenario* s_0 *is the one that is getting impossible, as we have seen. Conditioning on* $C_A(\xi \wedge in \wedge out) = S_A - \{s_0\}$, *we get the posterior probabilities* $p(s_0) = 0$ *and* $p(s_1) = 0.315$ *for example.*

These probabilities represent added value to the inference. For example, it is often of interest to have the probability that a given assumption is positive (or negative). There are many ways that these probabilities give additional insight into the inference problem. As an illustration we are going to discuss a special kind of explanations (or diagnosis) introduced by Reiter [19]. 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 scenarios, we write $s' \leq s$, if $s_- \subseteq s'_-$. Reiter considers explanations which are minimal relative to this partial order in $PS_A(\omega, \xi)$. This makes especially in diagnostics much sense: it singles explanations out, where the number of assumed faulty components (negative assumptions) is minimal.

We have the following simple result:

Lemma 3.4.1. *If for all assumptions* $a \in A$ *we have that* $p(a) \geq p(\neg a) = 1 - p(a)$, *then for* $s, s' \in PS_A(\omega, \xi)$ *we have that* $s' \leq s$ *implies* $p(s') \leq p(s)$.

Proof. Assume without loss of generality that $s = \neg a_1 \wedge \cdots \wedge \neg a_l \wedge a_{l+1} \wedge \cdots \wedge a_m$, $s = \neg a_1 \wedge \cdots \wedge \neg a_k \wedge a_{k+1} \wedge \cdots \wedge a_m$, 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:

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

Last equation holds, because $1 - p(a_i) \leq p(a_i)$, $i = 1, \ldots, m$, and the the lemma is proven. □

The assumption that $p(a) \geq p(\neg a)$ is in diagnostics again 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 scenarios in their downsets $s^{\downarrow} = \{s' \in PS_A(\omega, \xi) | s' \leq s\}$.

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

$$\begin{aligned}
p(\tilde{s}|C_A(\xi \wedge \gamma \wedge \tau)) &= \max_{s \in C_A(\xi \wedge \gamma \wedge \tau)} p(s|C_A(\xi \wedge \gamma \wedge \tau)) \\
&= k \cdot \max_{s \in C_A(\xi \wedge \gamma \wedge \tau)} p(s). \qquad (12)
\end{aligned}$$

Note that there may be several maximum likelihood scenarios. It follows immediately, that any maximum likelihood scenario is a Reiter explanation (but not the inverse).

Example 3.4.2. *Consider Example 3.4.1. The maximum likelihood scenario 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.*

4 Conclusion

In this paper we examined in particular abduction and logic-based diagnostics. Probabilistic argumentation systems are 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 scenarios and to delimit thus the remaining possible scenarios. 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 thus 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.

Probabilistic argumentation systems combine 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 criticise 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. However this has to be worked out yet.

References

[1] B. Anrig, R. Bissig, R. Haenni, J. Kohlas, and N. Lehmann, *Probabilistic argumentation systems: Introduction to assumption-based modeling with ABEL*, Tech. Report 99-1, University of Fribourg, Institute of Informatics, Theoretical Computer Science, 1999.

[2] A. Darwiche, *Model-based diagnosis using structured system descriptions*, Artificial Intelligence Research **8** (1998), 165–222.

[3] ———, *Model-based diagnosis under real world constraints*, (2000), To appear in AI Magazine.

[4] D.T. Dupré, *Abductive and consistency-based diagnosis revisited: a modeling perspective*, Proc. of the 8th International Workshop on Non-Monotonic Reasoning, NMR'2000, 2000.

[5] R. Haenni, *Modeling uncertainty with propositional assumption-based systems*, Applications of Uncertainty Formalisms (S. Parson and A. Hunter, eds.), Lecture Notes in Artifical Intelligence 1455, Springer, 1998, pp. 446–470.

[6] R. Haenni, J. Kohlas, and N. Lehmann, *Probabilistic argumentation systems*, Tech. Report 99-9, University of Fribourg, Institute of Informatics, Theoretical Computer Science, 1999.

[7] ———, *Probabilistic argumentation systems*, Defeasible Reasoning and Uncertainty Management Systems: Algorithms (J. Kohlas and S. Moral, eds.), Oxford University Press, 2000.

[8] R. Haenni and N. Lehmann, *Buidling argumentation systems on set constraint logic*, Information, Uncertainty and Fusion (B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, eds.), Kluwer Academic Publishers, 2000, pp. 393–406.

[9] A. Kean, *The approximation of implicates and explanations*, International Journal of Approximate Reasoning **9** (1993), 97–128.

[10] J. Kohlas, B. Anrig, R. Haenni, and P.A. Monney, *Model-based diagnostics and probabilistic assumption-based reasoning*, Aritficial Intelligence **104** (1998), 71–106.

[11] J. Kohlas and P.A. Monney, *A mathematical theory of hints. An approach to the dempster-shafer theory of evidence*, Lecture Notes in Economics and Mathematical Systems, vol. 425, Springer, 1995.

[12] K.B. Laskey and P.E. Lehner, *Assumptions, beliefs and probabilities*, Artificial Intelligence **41** (1989), 65–77.

[13] E. Neufeld and D. Poole, *Combining logic and probability*, Comp. Intelligence **4** (1988), 98–99.

[14] J. Pearl, *Probabilistic reosoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann Publishers, Inc., 1988.

[15] D. Poole, *A logical framework for default reasoning*, Artificial Intelligence **36** (1988), 27–47.

[16] ———, *Representing knowledge for logic-based diagnosis*, Proc. International Conference of Fifth Generation Computer Systems, 1988, pp. 1282–1290.

[17] ______, *Normality and faults in logic-based diagnosis*, Proc. Eleventh International Joint Conference on Artificial Intelligence, 1989, pp. 1304–1310.

[18] G.M. Provan, *A logic-based analysis of dempster-shafer theory*, Int. J. Approximate Reasoning **4** (1990), 451–495.

[19] R. Reiter, *A theory of diagnosis from first principles*, Artificial Intelligence **32** (1987), 57–95.

[20] G. Shafer, *The mathematical theory of evidence*, Princeton University Press, 1976.

[21] Ph. Smets, *The transferable belief model for quantified belief representation*, Handbook of Defeasible Reasoning and Uncertainty Management Systems (D.M. Gabbay and Ph. Smets, eds.), vol. 1, Kluwer Academic Publishers, 1998, pp. 267–301.

[22] N. Wilson, *Algorithms for Dempster-Shafer theory*, Algorithms for Uncertainty and Defeasible Reasoning (J. Kohlas and S. Moral, eds.), Kluwer Academic Publishers, 1999.