

Abduction of distributed theories through local interactions

Gauvain Bourgne¹ and Katsumi Inoue¹ and Nicolas Maudet²

Abstract. What happens when distributed sources of information (agents) hold and acquire information locally, and have to communicate with neighbouring agents in order to refine their hypothesis regarding the actual global state of this environment? This question occurs when it is not be possible (*e. g.* for practical or privacy concerns) to collect observations and knowledge, and centrally compute the resulting theory. In this paper, we assume that agents are equipped with full clausal theories and individually face abductive tasks, in a globally consistent environment. We adopt a learner/critic approach. Previous work in this line mostly relied on some assumptions of compositionality (which allow to treat each piece of exchanged information separately). Because no shared background knowledge is assumed to start with, this does not hold here. We design a protocol guaranteeing convergence to a situation “sufficiently” satisfying as far as consistency of the system is concerned, and discuss its other properties.

1 Introduction

Many challenging questions in multiagent systems stems from the fact that the autonomous entities (agents) are inherently distributed, and that they perceive and act locally with little possibility of coordination with other agents. Often the system indeed comes with severe communication restrictions, due to physical (*e. g.* the limited scope of a communication device) or reasoning (*e. g.* the mere impossibility to consider all the potential communications) limitations of agents populating it. In such situations, a relevant issue to address is to understand to what extent these constraints affect the global behaviour of the system. In this paper we address this question in the specific context of agents performing hypothetical reasoning via abduction.

More precisely, the problem is as follows: a number of agents make local observations in a world (that we suppose globally consistent, *i. e.* if all the agents were together in big room to merge their informations they would not derive a contradiction). The objective of agents is to come up with a satisfying hypothesis with respect to their local perception of the world. The objective of the designer is to set up the system in such a way that the hypotheses reached by agents individually satisfy some sort of global property, despite the aforementioned communication constraints, without requiring agents to disclose their complete theories. Specifically, agents will be assumed to communicate only by means of bilateral interactions with a set of neighbours. The basic idea of the protocol used here is to let agents communicate when there is a good reason to do so, thus ensuring that communications are “relevant”. Of course by doing so we run into a difficult problem, for it is certainly not easy for agents to fore-

see which information will turn out to be relevant for other agents once passed to them.

There are different ways to circumvent this problem. The first is to make strong assumptions on the underlying theories held by agents. One such assumption is that the consistency property is *compositional*. Intuitively, this helps because elementary observations can then be considered independently, allowing simple protocols to be designed. For instance, under this assumption, it suffices to check that an hypothesis is consistent with two agents separately to conclude that it still is with the theories of both agents taken together. This is not realistic: often agents will have to join theories to find out that a given hypothesis is not consistent. In this paper we do not make such an assumption. Agents represent their knowledge with full clausal theories, *i. e.* sets of first order clauses, which is an expressive subset of first order logic, for which sound and complete methods for abduction have been proposed in the centralized case [5].

Distributed abduction has been considered in recent years in the ALIAS system [4]. They distribute the abductive programming algorithm of [6], using abductive logic program to represent each agent’s theory. More recently, DARE [7] addressed a similar problem, but consider possible dynamicity of the system by allowing agents to enter or exit some proof cluster. In none of these works however is the issue of communication constraints explicitly raised. Another related work is the peer-to-peer consequence finding algorithm DeCA [1]. Based on a different method (splitting clauses), it is to our knowledge the only other work in this domain taking into account restrictions of communication between peers. It is however restricted to propositional theories. Finally, our work is related to the approach of [2]. However the protocols obtained in this work mostly rely on an assumption that will not hold in our context, as we shall see below. This paper thus advances the state of the art by coping simultaneously with these two difficulties: the expressivity of the language used to represent agents’ theories, and the restrictions imposed on the communication between agents.

The rest of this paper is as follows. Section 2 gives the necessary background on abduction and consequence finding. The general learner/critic approach is presented in Section 3, where we also present the basic assumptions of our system. The core of the paper lies in Section 4 where the protocol for agents to locally interact is described and illustrated by examples. The main properties of our protocol are then given, and finally, Section 5 concludes.

2 Abduction and consequence finding

2.1 Preliminaries

First, we review some notions and terminology to represent our problem in a logical setting. A *literal* is an atom or the negation of an

¹ National Institute of Informatics, Tokyo, Japan. {bourgne,ki}@nii.ac.jp

² LAMSADE, Univ. Paris Dauphine, France. nicolas.maudet@dauphine.fr

atom. A *clause* is a disjunction of literals, and is often denoted by the set of literals. A clause $\{A_1, \dots, A_m, \neg B_1, \dots, \neg B_n\}$, where A_i and $\neg B_j$ are respectively positive and negative atoms is also written as $A_1 \vee \dots \vee A_m \leftarrow B_1 \wedge \dots \wedge B_n$. Any variable in a clause is assumed to be universally quantified at the front. A *clausal theory* is a finite set of clauses which can be identified with the conjunction of the clauses.

Let S and T be clausal theories. S *logically implies* T , denoted as $S \models T$, if and only if for every interpretation I such that S is true under I , T is also true under I . $\models$ is called the *entailment relation*. For a clausal theory T , a *consequence* of T is a clause entailed by T . We denote by $Th(T)$ the set of all consequences of T .

We can now introduce the notion of *characteristic clauses*, which represents “interesting” consequences of a given problem [5]. Each characteristic clause is constructed over a sub-vocabulary of the representation language called a *production field*, and represented as $\langle \mathcal{L} \rangle$, where $\mathcal{L}$ is a set of literals closed under instantiation. A clause C belongs to $\mathcal{P} = \langle \mathcal{L} \rangle$ if every literal in C belongs to $\mathcal{L}$. For a clausal theory T , the set of consequences of T belonging to $\mathcal{P}$ is denoted $Th_{\mathcal{P}}(T)$. Then, the characteristic clauses of T wrt to $\mathcal{P}$ are defined as $Carc(T, \mathcal{P}) = \mu Th_{\mathcal{P}}(T)$, where μ is subsumption minimality³.

When a set of new clauses S is added to a clausal theory, some consequences are newly derived with this additional information. The set of such clauses that belong to the production field are called *new characteristic clauses* of S wrt T and $\mathcal{P}$; they are defined as $Newcarc(T, S, \mathcal{P}) = Carc(T \cup S, \mathcal{P}) \setminus Carc(T, \mathcal{P})$.

2.2 Abductive hypothesis

The logical framework of hypothesis generation in abduction for the centralized case can be expressed as follows. Let T be a clausal theory, which represents the *background theory*, and O be a set of literals, which represents *observations*. Also let $\mathcal{A}$ be a set of literals representing the set of *abducibles*, which are candidate assumptions to be added to T for explaining O . Given T , O and $\mathcal{A}$, the abduction problem is to find a set H of literals such that:

- (i) $T \cup H \models O$ (accountability),
- (ii) $T \cup H \not\models \perp$ (consistency), and
- (iii) H is a set of instances of literals from $\mathcal{A}$ (bias).

In this case, H is also called an *explanation* of O (with respects to T and $\mathcal{A}$). An hypothesis is *minimal* if no proper subset of H satisfies the above three conditions (which is equivalent to subsumption minimality for ground clauses). An hypothesis is *ground* if it is a set of ground literals (literals containing no variable). This restriction is often employed in applications whose observations are also given as ground literals. In the following, we shall indeed assume that observations are grounded, and that we are only searching for minimal ground hypotheses.

2.3 Computation through hypothesis finding

Given the observations O , each hypothesis H of O can be computed by the principle of *inverse entailment* [5, 8], which converts the accountability condition (i) to $T \cup \{\neg O\} \models \neg H$, where $\neg O = \bigvee_{L \in O} \neg L$ and $\neg H = \bigvee_{L \in H} \neg L$. Note that both $\neg O$ and $\neg H$ are clauses since O and H are sets of literals. Similarly, consistency condition (ii) is equivalent to $T \not\models \neg H$. Hence, for any hypothesis H ,

³ meaning that μX represents the clauses of X that are not properly subsumed by any other clause of X .

its negated form $\neg H$ is deductively obtained as a “new” theorem of $T \cup \{\neg O\}$ that is not an “old” theorem of T alone. Moreover, to respect the bias condition (iii), every literal of $\neg H$ has to be an instance of a literal in $\bar{\mathcal{A}} = \{\neg L \mid L \in \mathcal{A}\}$. Then the negation of minimal hypotheses are the new characteristic clauses of O with respect to T and $\bar{\mathcal{A}}$, that is, $Newcarc(T, \{\neg O\}, \bar{\mathcal{A}})$.

SOLAR [9] is a sophisticated deductive reasoning system based on SOL-resolution [5], which is sound and complete for finding *minimal* consequences belonging to a given language bias (a *production field*). Consequence-finding by SOLAR is performed by *skipping* literals belonging to a production field $\mathcal{P}$ instead of resolving them. Those skipped literals are then collected at the end of a proof, which constitute a clause as a logical consequence of the axiom set. Using SOLAR, we can implement an abductive system that is *complete* for finding minimal explanations due to the completeness of consequence-finding.

Once possible hypotheses have been computed, a ranking process can be applied to select a *favorite hypothesis*. We shall here assume that a *preference relation* $\geq_p$ over the hypothesis is given as a total order between sets of grounded literals.

Example 1 We present here a problem that will be used to illustrate our method, using a full clausal theory with 10 clauses. $T = \{b(X) \vee \neg g(X), \neg e(X) \vee \neg c(X), \neg a(c_1), b(X) \vee \neg k(X), e(X) \vee \neg h(X), f(X, Y) \vee \neg h(Y) \vee \neg g(X), k(X) \vee \neg i(X), d(X, Y) \vee \neg f(X, Y), a(X) \vee c(X) \vee \neg d(Y, X), h(c_1)\}$. Our problem is to find explanations of $O = \{b(c_2)\}$ given the abducibles $\mathcal{A} = \{g(X), i(X)\}$. Then, with production field $\mathcal{P} = \{\neg g(X), \neg i(X)\}$, we can compute : $Carc(T \cup \{\neg b(c_2)\}, \mathcal{P}) = \{(\neg g(c_2)), (\neg i(c_2))\}$, and $Carc(T, \mathcal{P}) = \{\neg g(c_2)\}$, so $Newcarc(T, \neg b(c_2), \mathcal{P}) = \{\neg i(c_2)\}$. The only explanation is thus $h = i(c_2)$.

3 Learner/critic approach

To deal with distributed hypothesis formation in multi-agent systems, we take a learner-critic approach, in which learner agents aim at producing a globally adequate hypothesis through internal computations and local interactions with other agents acting as critics.

System. We consider in the remainder of this paper a system of n agents $a_0, \dots, a_{n-1}$. They share a common language $\mathcal{L}$ as well as a common set of abducibles $\mathcal{A}$ and a common preference relation $\geq_p$ over hypotheses. However, each agent a_i has its own *individual theory* T_i and its own *observations* O_i . Agent i will also form its own favorite hypothesis h_i , though it can adopt the hypothesis of other agents. In fact, at the end of the process, all agents will share the same hypothesis. Theories and observations are considered as certain. As such, they are assumed to be consistent, meaning that $\bigcup_{i < n} T_i \cup \bigcup_{i < n} O_i \not\models \perp$. To ensure termination, it will also be assumed that $Carc(\bigcup_{i < n} T_i, \langle \mathcal{L} \rangle)$ is finite, and that both hypotheses and observations are grounded (i.e. contain no variable). Moreover, we consider *communicational constraints* imposing that each agent can only communicate with a fixed set of neighbours.

Explainability. Our objective is then to ensure the *explainability* of the hypothesis with certain knowledge, at different levels. In our abductive setting, a set of grounded literals h will be said to be an explanation of (T, O) if it is consistent and complete wrt T and O (see conditions (i) and (ii) defined above), which is will be denoted by

$Expl(h, (\mathcal{T}, O))$. Thus, an hypothesis h is an abductive explanation of O with respect to $\mathcal{T}$ and $\mathcal{A}$ if it is a set of grounded literals of $\mathcal{A}$ such that $Expl(h, (\mathcal{T}, O))$ is true.

We now extend this notion of explainability to agents and group of agents. The basic notion that we shall use is a notion a *group explainability*. An agent a_i will be said to have a group explanation wrt. the group of agents G ($GExpl(a_i, G)$) iff $Expl(h_i, (\cup_{i \in G} \mathcal{T}_i, \cup_{i \in G} O_i))$.

Partial hypotheses. In order for a learner agent to propose an hypothesis to a critic, this agent must be able to produce such an hypothesis. However, given only a few clauses of the whole clausal theory, it might not be possible to find an explanation for the observations using only abducibles. Therefore, we shall allow an agent to build *partial hypotheses* which contain some non-abducible literals. Those literals might be the unexplained observations, or preferably some other literals of the language that would explain it. While interacting with other agents, they will share knowledge to expand these hypotheses in order to progressively build a fully abducible one. Note that of course, an hypothesis respecting the bias condition will always be favored over one which does not.

Lack of compositional explainability. Previous works using a learner/critic approach relied heavily on a simplifying assumption allowing agents to check the hypotheses on a strictly local basis, the *compositional explainability*. It is defined as the equivalence between (i) $Expl(h, (\mathcal{T}_i, O_i)) \wedge Expl(h, (\mathcal{T}_j, O_j))$, and (ii) $Expl(h, (\mathcal{T}_i \cup \mathcal{T}_j, O_i \cup O_j))$. Unfortunately, abduction with full clausal theories fails to meet this property. Still, some properties are preserved (at least as long as we assume, as we do here, that the underlying entailment relation is classical). Specifically, consistency enjoys *incrementality* ((i) $\Leftarrow$ (ii)), which ensures that non explainability is monotonic) and accountability enjoys *additivity* ((i) $\Rightarrow$ (ii)). This means that although consistency cannot be checked by merely performing individual evaluation of a given candidate hypothesis, the fact that a counter-example can be found is sufficient to dismiss it. On the other hand, regarding accountability checking, it may be that what an agent considers as a counter-example turns out not to be one, so the learner should be able to “argue” as to why he disagrees.

4 Protocol for distributed abduction

We describe here our protocol for distributed abduction. It follows a learner/critic approach, with a learner agent computing locally its hypothesis before interacting with a critic agent. However, since explainability is not compositional, the critic agent cannot ensure the consistency of the hypothesis by himself. Thus, a context will be incrementally built by using in turn the theories of both agents to check any possible incoherence. Moreover, since an agent might not be able to compute locally an hypothesis accountable for all observations, we allow the proposal of partial hypotheses. As the critic agent might have better grounds to build a good hypothesis, we will also allow it to reverse the roles by acting in turn as learner. We first describe the interaction of two agents, before giving more details on the consistency check. Then we shall describe the behaviour with n agents, and prove the properties of this mechanism.

4.1 Local interaction

An interaction is initiated by a learner agent a_0 proposing its hypothesis and its validity context to a critic agent a_1 . If the pro-

posed hypothesis h_0 is a new one, the first context Ctx_0 will be computed as the new consequences of $h_0 \cup \mathcal{T}_0$ wrt to h_0 , that is $Newcnc(\mathcal{T}_0, h_0, \mathcal{P}_L)$ where $\mathcal{P}_L = \langle \mathcal{L} \rangle$. Otherwise, the previously computed context will be used as the initial context Ctx_0 . Then, upon receiving such a proposal, a_1 will start its critic, which consists of several steps:

1. *Consistency check.* First, the critic agent checks that the proposed context and hypothesis are consistent with its own theory, before augmenting the context with conclusions from its own theory, and submitting it back to the learner. Through exchange between both agents, the context is updated to contain every new consequences of $h_0 \cup \mathcal{T}_0 \cup \mathcal{T}_1$. This process is further detailed in the next subsection. It can terminate in two ways.
 - If an inconsistency is detected, both agents rule out the hypothesis by adding its negation to their theories. The learner agent then moves on to its next preferred hypothesis and proposes it, triggering a new critic phase.
 - Otherwise, both agents memorize the final context Ctx_f of this hypothesis, and the critic phase moves to the next step.
2. *Accountability check.* In this step, the critic agent check if all its observations are explained by $h_0 \cup \mathcal{T}_1$.
 - If an unexplained observation o is found, it is sent to the learner agent. We then have two possibilities.
 - If o is not explained by $h_0 \cup \mathcal{T}_0$, it is a true counter-example. The learner agent then computes a new hypothesis that will also cover o , and propose it, triggering a new critic phase.
 - If o is already explained by $h_0 \cup \mathcal{T}_0$, then the learner agent will notify the critic of this fact and indicate which part p_0 of the hypothesis is used in explaining o with $\mathcal{T}_0$. The critic agent will add the clause $\{o \vee \neg p_0\}$ in its theory⁴. This new information will ensure that the critic agent can find the hypothesis on its own in further steps, or build up upon it. It will then proceed to the next unexplained observation.
 - If there is no unexplained observation, the critic proceeds to the next step.
3. *Acceptability check.* Any hypothesis reaching this step is consistent and accounts for the observations, but might include some non-abducible literals, or unnecessary parts. This step ensures that alternative hypotheses are explored if needed.
 - If the critic holds an hypothesis h_c that is preferred to h_0 (according to $\geq_p$), it will reverse roles and submit it. This will finally either result in the acceptance of a better hypothesis, or cause the former critic agent to learn why its hypothesis cannot be used.
 - Otherwise, if the hypothesis contains non-abducibles, the critic agent will try to compute a new one from the data it gathered and propose it. If it cannot find any other hypothesis, it will ask the learner agent to do so. Either way, h_0 will be temporarily ruled out by both agents. If none of them can propose a new hypothesis, then the interaction will settle on one of the previously proposed ones, moving to the next step.
 - If the hypothesis is acceptable, then the critic moves to the acceptance step.
4. *Acceptation.* In this step, the final outcome of the interaction is chosen. Hypotheses that were temporarily ruled out are restored,

⁴ Note that since p is a conjunction of literals, $\neg p_0$ is indeed a clause.

and the best hypothesis is chosen as the final hypothesis. It is adopted and memorized by both agents, ending the interaction.

Example 2 We consider the same theory as in the last example, but now distributed among two agents.

$$\begin{aligned} \mathcal{T}_0 &= \{ b(X) \vee \neg g(X), \\ &\neg a(c_1), b(X) \vee \neg k(X), \\ &e(X) \vee \neg h(X), \\ &f(X, Y) \vee \neg h(Y) \vee \neg g(X) \} \\ \mathcal{T}_1 &= \{ \neg e(X) \vee \neg c(X), \\ &a(X) \vee c(X) \vee \neg d(Y, X), \\ &k(X) \vee \neg i(X), h(c_1), \\ &d(X, Y) \vee \neg f(X, Y) \} \end{aligned}$$

Moreover, $O_0 = \emptyset$ and $O_1 = \{b(c_2)\}$. The language of abducible remains the same. Initially, a_0 has hypothesis $h_0^0 = \emptyset$ and a_1 has hypothesis $h_1^0 = b(c_2)$ as it has no clause containing $b(X)$.

If agent a_0 starts the interaction, it first proposes its empty hypothesis, which goes directly through the consistency check of a_1 since it has no new consequences. Then in the accountability step, a_1 sends $b(c_2)$ as a counter example, and a_0 has to compute a new hypothesis. It gets $h_0^1 = g(c_2)$, which will be proved inconsistent in the consistency check (details in next section), resulting in the computation of a new hypothesis, $h_0^2 = k(c_2)$, who goes through the consistency check. Then, during accountability check, a_1 considers that $b(c_2)$ is not explained by $k(c_2)$ and sends it as a counter-example. Since $\mathcal{T}_0$ can explain this observation with $k(c_2)$, it returns $(b(c_2) \vee \neg k(c_2))$ as an argument (note that the rule is then instantiated). This rule is added to the theory of a_1 , and as there are no other observations, a_1 moves to the acceptability check. Since $k(c_2)$ is not an abducible, it will not be accepted before exploring other hypothesis. h_1^0 is not better than h_0^2 , but since a_1 got a new rule, it can use it to compute new hypothesis. As a result it gets $h_1^1 = i(c_2)$, which respects the bias condition. Thus h_0^2 is temporarily denied and ruled out, before a_1 proposes $i(c_2)$ with the newly computed context $\{i(c_2), k(c_2)\}$. The context is confirmed, but during the accountability check, a_0 sends $b(c_2)$ as a counter-example. a_1 thus argues $(b(c_2) \vee \neg i(c_2))$ (grounding and combining its clauses), which is added to the theory of a_0 . Then, during acceptability step, a_0 cannot compute a better hypothesis than h_1^1 . This hypothesis is composed of abducibles, so there is no need to explore further, and it is accepted, ending the exchange. Note that at the end of this exchange, the only correct hypothesis has been found. Moreover, one rule has been added to $\mathcal{T}_0$, one other rule to $\mathcal{T}_1$, and observation $b(c_2)$ has been added to O_0 .

4.2 Consistency check through context computation

During consistency check, a validity context is incrementally build for a given hypothesis h to compute the new consequences of $h \cup \mathcal{T}_0 \cup \mathcal{T}_1$ wrt to $\mathcal{T}_0 \cup \mathcal{T}_1$. If the hypothesis is incoherent, then it will have $\perp$ as a consequence. The occurrence of a contradiction between the context and the agents' theories will thus enable detection of incoherent hypotheses. This relies on the fact that the global theory itself is assumed to be consistent, so any inconsistency can only arise from the hypotheses. The process is as follows:

1. The learner agent a_0 retrieves the context Ctx_0 of its hypothesis. If no context has been memorized from previous interaction, then a new one is computed as $Newarc(\mathcal{T}_0, h_0, \mathcal{P}_L)$ (note that we should thus have $h_0 \in Ctx_0$).
2. Upon receiving h_0 and Ctx_0 , agent a_1 first checks if it already has some context Ctx' for this given hypothesis. If it is the case,

then it replaces Ctx_0 by $Ctx_0 \cup Ctx'$. Context Ctx_1 is then computed as $Newarc(\mathcal{T}_1, Ctx_0, \mathcal{P}_L)$, and sent back to a_0 (unless a contradiction is found).

3. The process continues. At each step Ctx_i is computed by agent a_α as the new consequences $Newarc(\mathcal{T}_\alpha, Ctx_{i-1}, \mathcal{P}_L)$, where $\alpha = 0$ if i is odd, and $\alpha = 1$ otherwise. This computation stops when either a contradiction is found or Ctx_i is included in either Ctx_{i-1} or Ctx_{i-2} , if which case all consequences have been computed.
4. If no contradiction were found, the final context is computed as $Ctx_f = Ctx_i \cap Ctx_{i-1}$ where i is the final step. Any element in respectively $Ctx_i \setminus Ctx_{i-1}$ and $Ctx_{i-2} \setminus Ctx_{i-1}$ are added to $\mathcal{T}_{1-\alpha}$ and $\mathcal{T}_\alpha$ where again $\alpha = 0$ if i is odd and 1 otherwise. Indeed an element will only be removed from the context if it is a direct consequence of one of the agent's theory.

Example 3 We will detail the consistency check of our last example for hypothesis $h_0^1 = \{g(c_2)\}$. When producing this hypothesis, since it is its first occurrence, a_0 computes Ctx_0 and proposes it with h_0^1 . Then a_1 computes Ctx_1 , and a_0 in turn can compute Ctx_2 . We thus have the following, with $\mathcal{P}_L = \langle \mathcal{L} \rangle$:

$$\begin{aligned} Ctx_0 &= Newarc(\mathcal{T}_0, h_0^1, \mathcal{P}_L) = \{g(c_2), f(c_2, X) \vee \neg h(X)\} \\ Ctx_1 &= Newarc(\mathcal{T}_1, Ctx_0, \mathcal{P}_L) = Ctx_0 \cup \{d(c_2, X) \vee \neg h(X), \\ &f(c_2, c_1), d(c_2, c_1), a(X) \vee \neg e(X) \vee \neg h(X), a(c_1) \vee \neg e(c_1), \\ &c(X) \vee a(X) \vee \neg h(X), c(c_1) \vee a(c_1)\} \end{aligned}$$

$$Ctx_2 = Newarc(\mathcal{T}_0, Ctx_1, \mathcal{P}_L) = Ctx_0 \cup \{d(c_2, X) \vee \neg h(X), f(c_2, c_1), d(c_2, c_1), a(X) \vee \neg h(X), c(c_1), \neg e(c_1), \neg h(c_1)\}$$

Note that between Ctx_1 and Ctx_2 , some clauses of Ctx_1 have been replaced by new clauses that subsumed them. When a_1 receives Ctx_2 , an inconsistency can easily be derived between $h(c_1) \in \mathcal{T}_1$ and $\neg h(c_1) \in Ctx_2$. The hypothesis is thus rejected.

4.3 System of n agents

Each interaction allows the participants to refine their hypotheses and augment their knowledge concerning their consequences. The protocol described before is sufficient to allow two agents to come up with an hypothesis that is coherent with the union of their theories, and it ensures that an hypothesis respecting the bias and accounting for all observations will be found if it can be done with the union of their theories. When more agents are involved, it is possible to chain such interactions to converge towards a consistent state of the system. To take into account potentially dynamic communication constraints in the system, we propose a rumor-like approach, ensuring that despite the local behaviour and interactions of the agents, the system converges to a state in which all agents have a group explanation with respect to the whole system (meaning that $\forall i, GExpl(a_i, \{a_0, \dots, a_{n-1}\})$). Each agent has the goal to ensure it has an explanation with respect to its neighbours, and it will seek to interact with them whenever required to ensure this. In practice, an agent a_i will engage in a local interaction with a neighbour a_j whenever its hypothesis and context (h_i, Ctx^i) differ from those obtained during its last interaction with a_j .

Example 4 We consider 3 agents a_0, a_1, a_2 such that a_0 and a_2 cannot communicate together, but both can communicate with a_1 . Their local theory are as follows, with $O_0 = O_1 = \emptyset$ and $O_2 = \{b(c_2)\}$:

$$\begin{aligned} \mathcal{T}_0 &= \{ \neg e(X) \vee \neg c(X), b(X) \vee \neg k(X), h(c_1) \} \\ \mathcal{T}_1 &= \{ e(X) \vee \neg h(X), f(X, Y) \vee \neg h(Y) \vee \neg g(X), b(X) \vee \neg g(X) \} \\ \mathcal{T}_2 &= \{ d(X, Y) \vee \neg f(X, Y), a(X) \vee c(X) \vee \neg d(Y, X), \neg a(c_1), k(X) \vee \neg i(X) \} \end{aligned}$$

First a_1 proposes $\emptyset$ to a_2 and receives counter example $b(c_2)$ in return. It then proposes $g(c_2)$. After consistency check, both agents agree on $Ctx_f^1 = \{g(c_2), f(c_2, X) \vee \neg h(X), d(c_2, X) \vee \neg h(X), a(X) \vee c(X) \vee \neg h(X), c(c_1) \vee \neg h(c_1)\}$, which is accepted. Then a_1 interacts with a_0 , and an incoherence is found for $g(c_2)$ (using Ctx_f^1 as Ctx_0 to initiate consistency check). Both agents add $\neg g(c_2)$ to their theory and then a_1 proposes $b(c_2)$ which is temporarily ruled out. Then a_0 gets the observation when proposing $\emptyset$, and recomputes its hypothesis to propose $k(c_2)$. It argues for it during accountability check, so a_1 adds the rule $(b(c_2) \vee \neg k(c_2))$. Since k is not an abducible, and a_1 has no other hypothesis, it asks for another one from a_0 . But a_0 has also exhausted its options, so finally the agents settle on $k(c_2)$ (with context $k(c_2)$). The third interaction implies again a_1 and a_2 : a_1 proposes $k(c_2)$ and argues for it (increasing T_2), but in acceptability check, a_2 counter proposes with $g(c_2)$, which is refuted by the rule $(\neg g(c_2))$ in T_1 . Then, using the rule $(b(c_2) \vee \neg k(c_2))$ it has just learned, a_2 can compute and propose $i(c_2)$, with context $\{i(c_2), k(c_2)\}$, which is accepted once a_2 has argued its accountability. Finally, a_1 proposes $i(c_2)$ to a_0 , and argues its accountability. The hypothesis and context are then accepted, and since all agents share the same hypothesis and context, the process stops.

4.4 Soundness of the protocol

Lemma 1 *We consider a consistency check between a_0 and a_1 with respect to hypothesis h , initiated by a_0 . Let c be a clause. If $\exists i, c \in Th_{\langle L \rangle}(Ctx_i)$ and $c \notin Th_{\langle L \rangle}(T_0 \cup T_1)$, then $\forall j > i$ we also have $c \in Th_{\langle L \rangle}(Ctx_j)$. In other words, any consequence of a context that is not a consequence of the theories remains a consequence of all subsequent contexts.*

Proof If $c \in Th_{\langle L \rangle}(Ctx_i)$, then c is a consequence of $T_\alpha \cup Ctx_i$ for any α . It means that, with $i + 1 \equiv \alpha[2]$, c is in $Th_{\langle L \rangle}(T_\alpha \cup Ctx_0) \setminus Th_{\langle L \rangle}(T_\alpha) = Th_{\langle L \rangle}(Ctx_{i+1})$ (given that $c \notin Th_{\langle L \rangle}(T_0 \cup T_1)$). By induction, it can thus be proved for any $j > i$. $\square$

Lemma 2 (n agents) *We consider a series of interactions I^k between different agents in a system, where I^k is an interaction between agents a_0^k and a_1^k such that there is a common agent between any successive interaction ($\forall k, \exists \alpha, \beta$ s. t. $a_\alpha^k = a_\beta^{k+1}$). During these interactions we consider the contexts generated for hypothesis h (proposed by a_0^k), denoting by Ctx_i^k the context number i of the interaction I^k . Let $c \notin Th_{\langle L \rangle}(h)$ be a new consequence of h (possibly contradictory) wrt $\bigcup_{i < n} T_i$. Then, there exists k_0 such that:*

- (a) $\exists q_0, \forall p > q_0, c \in Th_{\langle L \rangle}(Ctx_p^{k_0})$;
- (b) $\forall l > k_0, \forall r > 0, c \in Th_{\langle L \rangle}(Ctx_r^l)$

Proof If c is a consequence of $\bigcup_{i < n} T_i \cup h$ that is not a consequence of $\bigcup_{i < n} T_i$, then there exists a proof from h , that is, a sequence of inference steps $s_0, \dots, s_{m-1}$ with $s_k = (c_k, d_k, c_{k+1})$ and $m > 0$ such that (i) $c_0 \in h$, (ii) $c_m = c$, (iii) $\forall k, d_k \in \bigcup_{i < n} T_i \cup h$, (iv) $\forall k, c_k \notin Th_{\langle L \rangle}(\bigcup_{i < n} T_i)$, (v) $\forall k, d_k \wedge c_k \models c_{k+1}$.

We shall then prove our lemma by induction on the number of inference step of the proof m . We first prove that $\exists k_0 \exists q_0, c \in Th_{\langle L \rangle}(Ctx_p^{k_0})$. If $m = 0$, we directly have $c \in h \setminus Th_{\langle L \rangle}(\bigcup_{i < n} T_i)$ implies $c \in Th_{\langle L \rangle}(Ctx_0^0)$ ($k_0 = q_0 = 0$). If $m > 0$, we assume that the lemma is true for c_{m-1} . Thus, $\exists k_0$ such as (a) and (b) are true with respect to c_{m-1} . The last step to prove c is $s_{m-1} = (c_{m-1}, d_{m-1}, c_m)$. if $d_{m-1} \in h$, it will be in $Th_{\langle L \rangle}(Ctx_0^{k_0})$ and

thus both d_{m-1} and c_{m-1} are consequences of $Ctx_{q_0+1}^{k_0}$ which ensures the same for c . Thus, given that $c \notin Th_{\langle L \rangle}(\bigcup_i T_i)$, $c \in Th_{\langle L \rangle}(Ctx_{q_0+1}^{k_0})$. Otherwise, d_{m-1} is in T_γ . There exists k'_0 such that a_γ is involved in $I^{k'_0}$ (otherwise, the interactions would continue as context would be different between a_γ and $a_0^{k_0}$). If $k'_0 = k_0$, (a) give q_0 s. t. $\forall p > q_0, c_{m-1} \in Th_{\langle L \rangle}(Ctx_p^{k_0})$. We can chose $p = q'$ such that $Ctx_{q'}^{k_0}$ is produced by a_γ , then $T_\gamma \cup Ctx_{q'-1}^{k_0}$ implies d_{m-1} and c_{m-1} (as $c_{m-1} \in Ctx_{q'}^{k_0}$), which means it implies $c_m = c$. This in turn, given that $c_m \notin \bigcup_{i < n} T_i$, ensures $c \in Th_{\langle L \rangle}(Ctx_{q'}^{k'_0})$. Otherwise, we can chose $p = q'$ such that $Ctx_{q'-1}^{k_0}$ is produced by a_γ . Then (b) ensures that $c_{m-1} \in Th_{\langle L \rangle}(Ctx_{q'}^{k'_0})$ and the same reasoning gives $c \in Th_{\langle L \rangle}(Ctx_{q'}^{k'_0})$.

Then Lemma 1 ensures that c is also in $Th_{\langle L \rangle}(Ctx_p^{k'_0})$ for any $p > q'$, which gives property (a). About property (b), if $\forall r > 0, c \in Th_{\langle L \rangle}(Ctx_r^l)$ holds for a given l , then we can prove that it holds for $l + 1$ as follows. For interaction I^{l+1} , one of the agents involved has context Ctx' resulting from previous interaction, which entails c . If the agent having Ctx' is a_0^{l+1} , we have $Ctx_0^{l+1} = Ctx'$, and lemma 1 gives us $\forall r \geq 0, c \in Th_{\langle L \rangle}(Ctx_r^l)$. Otherwise, Ctx_1 is defined as the minimal elements of $Th_{\langle L \rangle}(Ctx_0 \cup Ctx' \cup T_\alpha) \setminus Th_{\langle L \rangle}(T_\alpha)$. Since c is a consequence of Ctx' (and thus also of $Ctx_0 \cup Ctx' \cup T_\alpha$) and is assumed not to be a consequence of T_α , we have $c \in Th_{\langle L \rangle}(Ctx_1^{l+1})$ and Lemma 1 ensures $\forall r > 0, c \in Th_{\langle L \rangle}(Ctx_r^l)$. By induction, it holds $\forall l > k_0$, proving (b). $\square$

Theorem 1 *Any hypothesis that has been accepted by all agents with the same context (computed through the consistency check procedure) is a group-explanation wrt the MAS.*

Proof By contradiction, if the hypothesis is inconsistent, then $\perp$ is a consequence of $h \cup \bigcup_{i < n} T_i$ that is not a consequence of $\bigcup_{i < n} T_i$ (since the theory is assumed to be consistent). Let I^0 be the first interaction in which h is accepted. If the system stabilizes with this hypothesis, it means that, starting from this interaction, we can construct a sequence of interactions I^k such that each interaction uses the context from the previous one. Then, applying Lemma 2, there exist k_0 and p_0 such that $\perp \in Th_{\langle L \rangle}(Ctx_{p_0}^{k_0})$. However, given the algorithm, in such case the inconsistency is found and the hypothesis discarded. Accountability can be checked locally as it is monotonic. So the system can only stabilize on a consistent hypothesis accounting for its observations, which means that the process is sound. $\square$

4.5 Termination

Theorem 2 (Termination of local interactions) *If hypotheses and observations are restricted to ground literals, and $Carc(\bigcup_{i < n} T_i)$ is finite, a local interaction will always terminate.*

Proof Given restrictions on the language and theory, the number of possible hypotheses and contexts that can be derived from the theory and the observations is finite. If the interaction does not terminate, it means that one of the step does not terminate. Since there is a finite number of possible hypotheses, and given that any hypothesis that reach acceptability step is ruled out, this step cannot repeat itself forever. Likewise, there is only a finite number of observations and a given observation can only be sent as a counter example once for an hypothesis, so completeness check also necessarily

terminates. What remains to be proved is that consistency check always terminates. By contradiction if it is not the case, then we have an infinite number of context Ctx_i , none of them being included in neither Ctx_{i-1} nor Ctx_{i-2} . Then let's consider the elements of $Ctx_i \setminus Ctx_{i-2}$ and $\alpha \in \{0, 1\}$ having the same parity as i . These elements are consequences of $h \cup T_0 \cup T_1$. If there are an infinite number of i such that these elements are new consequences of $h \cup T_0 \cup T_1$ wrt h , given Lemma 1, there would be an infinite number of consequences, which is not possible under our assumptions. Moreover, these elements cannot be in T_α since they belong to Ctx_i . Thus, $\exists i_0$ such that $Ctx_i \setminus Ctx_{i-2} = D \subseteq Th_{\langle L \rangle}(T_{1-\alpha})$. Then $Ctx_{i_0+1} = Th_{\langle L \rangle}(Ctx_{i_0} \cup T_{1-\alpha}) \setminus Th_{\langle L \rangle}(T_{1-\alpha}) = Th_{\langle L \rangle}(Ctx_{i_0-2} \cup D \cup T_{1-\alpha}) \setminus Th_{\langle L \rangle}(T_{1-\alpha}) = Th_{\langle L \rangle}(Ctx_{i_0-2} \cup T_{1-\alpha}) \setminus Th_{\langle L \rangle}(T_{1-\alpha})$, that is $Ctx_{i_0+1} = Ctx_{i_0-1}$ which ends the consistency check. $\square$

Theorem 3 (Convergence) *Given the individual behaviour of agent described before, the system will eventually reach a state in which every group of connected agents has a group-explanation wrt itself whenever termination of local interactions is ensured.*

Proof If there is an infinite number of local interaction, it means that at least one agent changes its hypothesis or its context an infinite number of times. But the only cases in which an agent changes its hypothesis is when either it is proved inconsistent or a better hypothesis is validated. In the latter case, the agent will only go back to a previous hypothesis if the one that replaced it is proved wrong, which can only happen a finite number of time. As for context, once a literal is part of the context of an agent, it will only be removed if a corresponding clause is included in its theory, so an agent will never change back to a context it had previously adopted. Thus, there cannot be an infinite number of interactions, which means that the system will reach a state in which all agents share the same hypothesis and context as their neighbours. $\square$

Note that this can be easily extended to system with dynamic communicational constraints if those constraints ensure that information can pass from one agent to another over time (which is ensured if at any time the connection graph of all present and future communication channel form a connected graph, a property called *temporal connectivity*). This is similar to the convergence through bilateral protocols presented for a simpler case of abduction with common theory in [3]. Moreover, if new pieces of informations or observations are given to the system at some time, the system will trigger necessary interactions to ensure the consistency of the hypotheses if given enough time: the described behaviour results in a *self-stabilizing* system.

4.6 Completeness

Since the system stabilizes on a single common hypothesis, completeness of the process (that is, all hypotheses that explain (T, O) are given as an output) is not ensured. However, we can guarantee that an hypothesis will always be produced if it is possible to do so.

Theorem 4 (Guarantee of finding an explanation) *If there exists at least one abductive explanation h of $\bigcup_{i < n} O_i$ wrt $\bigcup_{i < n} T_i$ and $\mathcal{A}$, and both termination of local interactions and temporal connectivity are guaranteed, then the MAS will eventually stabilize in a state in which all agents share such an explanation (i.e. a group-explanation wrt the MAS that respects the bias $\mathcal{A}$).*

Proof By contradiction, let us suppose the MAS stabilizes on a partial hypothesis. Let h be an abductive explanation, and let $s_0, \dots, s_{m-1}$ be a proof of $\neg h$ from $T \cup \{\neg O\}$ (with same notations $s_i = (c_i, d_i, c_{i+1})$ as in Lemma 1). $\exists k > 0, \exists \alpha$, s.t. $d_0, \dots, d_{k-1}$ are all in $T_\alpha \cup \{\neg O_\alpha\}$. If $k = m$, then a_α can compute h and will eventually propose it. Otherwise, assume by induction that $k = m - 1$. Then, $\exists \gamma$ such that $d_k \in T_\gamma \cup \{\neg O_\gamma\}$. We can then prove that h will eventually be proposed during the interaction of a_α and a_γ (as a_α should propose partial hypothesis $\neg c_k$ at some point, and a_γ would then get the rules to produce c_k by itself during accountability check, and will thus be able to compute h and propose it)⁵. Then, we can be sure that h will eventually be proposed, and given stabilization theorem, it will be adopted by all agents (unless they prefer another acceptable hypothesis). Either way, we get a contradiction. $\square$

When completeness is sought, the protocol should be modified to try every single candidate hypothesis before accepting one (which might be intractable in complex cases) and to stabilize on a set of hypotheses rather than on a unique one.

5 Conclusion

In this paper we investigated distributed abduction of full clausal theories with restricted communications. The problem is challenging because the explainability relation cannot be assumed to be compositional in this situation. Adopting a learner/critic approach, we proposed a learner/critic bilateral protocol which provably soundly terminates. A weak notion of completeness is also guaranteed, which is arguably what is required in most cases. Future works include considering specific distribution of the rules. Especially, if the size of the common language can be restricted, while each agent has a private language, then we only need to compute contexts pertaining to the common language, which should improve tractability. Moreover, since an agent only reveal information about the consequences of its theory, and not its internal reasoning, we can expect some interesting properties with respect to privacy.

REFERENCES

- [1] P. Adjiman, P. Chatalic, F. Goasdoué, M.-C. Rousset, and L. Simon. Distributed reasoning in a peer-to-peer setting: Application to the semantic web. *J. Artif. Intell. Res. (JAIR)*, 25:269–314, 2006.
- [2] G. Bourgne, A. El Fallah-Seghrouchni, and N. Maudet. Towards refinement of abductive or inductive hypotheses through propagation. *Journal of Applied Logic*, 7(3):289–306, 2009.
- [3] G. Bourgne, G. Hette, N. Maudet, and S. Pinson. Hypotheses refinement under topological communication constraints. In *Proc. of AAMAS'07*, pages 994–1001. ACM Press, May 2007.
- [4] A. Ciampolini, E. Lamma, P. Mello, F. Toni, and P. Torroni. Cooperation and competition in ALIAS: a logic framework for agents that negotiate. *Annals of Math. and Artif. Intell.*, 37(1–2):65–91, 2003.
- [5] K. Inoue. Linear resolution for consequence finding. *Artif. Intell.*, 56(2–3):301–353, 1992.
- [6] A. C. Kakas and P. Mancarella. Database updates through abduction. In *Proc. of VLDB '90*, pages 650–661. Morgan Kaufmann Pub., 1990.
- [7] J. Ma, A. Russo, K. Broda, and K. Clark. Dare: a system for distributed abductive reasoning. *JAAMAS*, 16(3):271–297, 2008.
- [8] S. Muggleton. Inverse entailment and progol. *New Generation Comput.*, 13(3&4):245–286, 1995.
- [9] H. Nabeshima, K. Iwanuma, and K. Inoue. Solar: A consequence finding system for advanced reasoning. In *Automated Reasoning with Analytic Tableaux and Related Methods*, pages 257–263. Springer, 2003.

⁵ Note that if a_α and a_γ are not connected, the rules to produce c_k will still be propagated by other agents until they reach a_γ .