INFSYS RESEARCH REPORT

INSTITUT FÜR INFORMATIONSSYSTEME

ABTEILUNG WISSENSBASIERTE SYSTEME

ON COMPUTING ALL ABDUCTIVE EXPLANATIONS

Thomas Eiter and Kazuhisa Makino

INFSYS RESEARCH REPORT 1843-02-04

APRIL 2002

Institut für Informationssysteme
Abtg. Wissensbasierte Systeme
Technische Universität Wien
Favoritenstraße 9-11
A-1040 Wien, Austria
Tel: +43-1-58801-18405
Fax: +43-1-58801-18493
sek@kr.tuwien.ac.at
www.kr.tuwien.ac.at

TECHNISCHE UNIVERSITÄT WIEN

INFSYS RESEARCH REPORT

INFSYS RESEARCH REPORT 1843-02-04, APRIL 2002

ON COMPUTING ALL ABDUCTIVE EXPLANATIONS

(PRELIMINARY REPORT)

Thomas Eiter[1] and Kazuhisa Makino[2]

Abstract. We consider the computation of all respectively a polynomial subset of the explanations of an abductive query from a Horn theory, and pay particular attention to whether the query is a positive or negative letter, the explanation is based on literals from an assumption set, and the Horn theory is represented in terms of formulas or characteristic models. We derive tractability results, one of which refutes a conjecture by Selman and Levesque, as well as intractability results, and furthermore also semi-tractability results in terms of solvability in quasi-polynomial time. Our results complement previous results in the literature, and elucidate the computational complexity of generating the set of explanations.

---

[1]Institut für Informationssysteme, Technische Universität Wien, Favoritenstraße 9-11, 1040 Wien, Austria. E-mail: eiter@kr.tuwien.ac.at.

[2]Division of Systems Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan. E-mail: makino@sys.es.osaka-u.ac.jp

Acknowledgements: This work was partly supported by the Austrian Science Fund (FWF) project Z29-INF, by TU Wien through a scientific collaboration grant, and by the Scientific Grant in Aid of the Ministry of Education, Science, Sports and Culture of Japan.

An abstract of this paper appears in: Proceedings 18th National Conference on Artificial Intelligence (AAAI '02), July 28 - August 1, 2002, Edmonton, Alberta, Canada.

Contents

1 Introduction

Abduction is a fundamental mode of reasoning, which has been recognized as an important principle of common-sense reasoning (see e.g. [3]). It has applications in many areas of AI including diagnosis, planning, learning, natural language understanding and many others (see e.g. references in [7]). In a logic-based setting, abduction can be defined as the task, given a set of formulas Σ (the background theory) and a formula χ (the *query*), to find a smallest set of formulas E (an *explanation*) from a set of hypotheses such that Σ plus E is satisfiable and logically entails χ.

For use in practice, the computation of abductive explanations is an important problem, for which well-known early systems such as Theorist [17] or ATMS solvers [4, 18] have been devised. Since then, there has been a vastly growing literature on this subject, indicating the need for efficient abductive procedures.

Main problems considered. In this paper, we consider computing a set of explanations for queries from Horn theories. More precisely, we address the following problems:

- Computing *all* explanations of a query χ given by a letter q, with and without a set of assumption literals A from which explanations E must be formed, similar as in [17, 20]. Note that the logical disjunction of all explanations is a weakest disjunctive form over the hypotheses explaining χ. It is easy to see that in general, there might be exponentially many explanations, and computing all explanations is inevitably exponential. However, it is in this case of interest whether the computation is possible in *polynomial total time* (or *output-polynomial time*), i.e., in time polynomial in the combined size of the input and the output. Furthermore, if exponential space is prohibitive, it is of interest to know whether a few explanations (e.g., polynomially many) can be generated in polynomial time, as studied by Selman and Levesque [20].

- We contrast formula-based (syntactic) with model-based (semantic) representation of Horn theories. The latter form of representation, where a Horn theory is represented by the *characteristic models*, was advocated by Kautz *et al.* [12]. As they showed, important inference problems are tractable in the model-based setting. Namely, whether a Horn theory Σ logically implies a CNF φ, and whether a query q has an explanation w.r.t. an assumption set A, as well as computing one; note that the latter is intractable under formula-based representation. Similar results were shown for other theories by Khardon and Roth [14].

- We investigate the role of syntax for computing abductive queries. In the framework of [20, 12], the query is a positive letter q. However, it is of equal interest to consider *negative queries* as well, i.e., to explain the complement $\overline{q}$ of an atom q. Since the Horn property imposes semantic restrictions on theories, it is not straightforward to express such negative queries in terms of positive queries.

- Finally, we consider as a meaningful generalization the computation of *joint explanations*. That is, given a background theory Σ and observations o_1, o_2,..., o_l, where $l \geq 2$, compute a *single* explanation E which is good for *each* o_i. Joint explanations are relevant, e.g., in diagnostic reasoning. We may want to know whether different observations allow to come up with the same diagnosis, given by an abductive explanation, about a system malfunctioning. Such a diagnosis is particularly strong, as it is backed up by several cases. Note that a joint explanation strengthens an explanation of the conjunction $o_1 \wedge o_2 \wedge \cdots \wedge o_l$ of observations.

Main results. Our main results are summarized as follows.

• We refute Selman and Leveque's belief [20, p. 266], that given a Horn theory Σ and a query letter q, it is hard to list all explanations of q from Σ even if we are *guaranteed* that there are only few explanations. More precisely, we disprove their conjecture that generating $O(n)$ many explanations of q is NP-hard, where n is the number of propositional letters in the language. This is a consequence of our result that generating *all* nontrivial explanations of q is possible in *total polynomial time* (Theorem 1).

• We give a detailed characterization of computing all explanations of a query from Horn theory in the formula- vs model-based setting, for both general explanations and explanations w.r.t. a set of assumption literals. In a nutshell, we obtain three kinds of results:

(1 A procedure which enumerates all nontrivial explanations of a query letter q from a Horn theory Σ with incremental polynomial delay. This is a positive result and trivially implies that all explanations can be found in polynomial total time. Moreover, it means that any polynomial number of explanations can be generated in polynomial time in the size of the input (Corollary 1).

(2) Intractability results for generating all explanations for a negative query $\overline{q}$ from a Horn theory Σ contained in a set of assumption literals A; this complements a similar result for positive queries in [20]. Both results emerge from the fact that the associated problems of recognizing the correct output are co-NP-complete. Since some hard instances have only small (polynomial-size) output, they also imply that computing few (polynomially many) explanations is intractable.

(3) Under model-based representation, generating all explanations is polynomial-time equivalent to the well-known problem of dualizing a positive CNF (DUALIZATION), i.e., given a CNF φ in which no negative literal occurs, compute the (unique) prime DNF of φ. DUALIZATION is a well-known open problem in NP-completeness, cf. [1, 10]; it is known to be solvable in *quasi-polynomial* total time, i.e., in time $N^{o(\log N)}$ where N denotes the combined size of the input and output [10]; furthermore, polynomial total time algorithms are known for many special cases, and as recently shown, the decisional variant of recognizing the prime DNF of φ is solvable with limited nondeterminism, i.e., in polynomial time with $O(\log^2 N)$ many guesses [8]. Since DUALIZATION is strongly believed not to be co-NP-hard, our result thus provides strong evidence that under model-based representation, computing all explanations is not co-NP-hard. Interestingly, the equivalence result holds for both positive and negative queries, and whether arbitrary or explanations over a set of assumption literals A are admitted. This means that, in a sense, model-based representation, in contrast to formula-based representation, is not sensitive to these aspects. Furthermore, by resorting to respective algorithms for dualization, the result provides us with algorithms for enumerating all or polynomially many explanations with quasi-polynomial time delay between outputs.

• We show that deciding the existence of a joint explanation is intractable, for both formula- and model-based representation. Thus, the positive results for ordinary explanations do not extend to joint explanations.

The rest of this paper is structured as follows. The next section contains preliminaries and fixes notation. In Section 3, we consider the task of generating all explanations for a query letter, and in Section 4 for a negated query letter. Section 5 is devoted to generating joint explanations. In Section 6, we review related work, and the final Section 7 concludes the paper. Technical details of some longer proofs have been moved to the Appendix.

2 Preliminaries

We assume a standard propositional language with letters $x_1, x_2, \ldots, x_n$ from a set P, where each x_i takes either value 1 (true) or 0 (false). Negated atoms are denoted by $\overline{x}_i$, and the opposite of a literal ℓ by $\overline{\ell}$. Furthermore, we use $\overline{A} = \{\overline{\ell} \mid \ell \in A\}$ for any set of literals A and set $Lit = P \cup \overline{P}$.

A clause is a disjunction $c = \bigvee_{p \in P(c)} p \vee \bigvee_{p \in N(c)} \overline{p}$ of literals, where $P(c)$ and $N(c)$ are the sets of atoms occurring positive and negated in c and $P(c) \cap N(c) = \emptyset$. Dually, a term is conjunction $t = \bigwedge_{p \in P(t)} p \wedge \bigwedge_{p \in N(t)} \overline{p}$ of literals, where $P(t)$ and $N(t)$ are similarly defined. We also view clauses and terms as sets of literals $P(c) \cup N(c)$ and $P(t) \cup N(t)$, respectively. A clause c is *Horn*, if $|P(c)| \leq 1$, and a CNF is *Horn*, if it contains only Horn clauses. A *theory* Σ is any set of formulas; it is *Horn*, if it is a set of Horn clauses. As usual, we identify Σ with $\varphi = \bigwedge_{c \in \Sigma} c$, and write $c \in \varphi$ etc.

Definition 1 Given a (Horn) theory Σ, called the background theory, a letter q (called query), and a set of literals $A \subseteq Lit$, an *explanation of* q *w.r.t.* A is a minimal set of literals E over A such that

(i) $\Sigma \cup E \models q$, and

(ii) $\Sigma \cup E$ is satisfiable.

If $A = Lit$, then we call E simply an *explanation of* q.

Observe that the above definition is slightly more general than the *assumption-based explanations* of [20], which emerge as $A = P' \cup \overline{P'}$ where $P' \subseteq P$ (i.e., A contains all literals over a subset P' of the letters). Furthermore, in some texts explanations must be sets of positive literals. As for Horn theories, the following is known (cf. [14]):

Proposition 1 *Let E be any explanation of q w.r.t. $A \subseteq Lit$. Then $E \subseteq P$, i.e., E contains only positive literals.*

Example 1 Consider a theory $\Sigma = \{\overline{x}_1 \vee \overline{x}_4, \overline{x}_4 \vee \overline{x}_3, \overline{x}_1 \vee x_2, \overline{x}_4 \vee \overline{x}_5 \vee x_1\}$. Suppose we want to explain $q = x_2$ from $A = \{x_1, x_4\}$. Then, we find that $E = \{x_1\}$ is an explanation. Indeed, $\Sigma \cup \{x_1\} \models x_2$, and $\Sigma \cup \{x_1\}$ is satisfiable; moreover, E is minimal. On the other hand, $E' = \{x_1, \overline{x}_4\}$ satisfies (i) and (ii) for $q = x_2$, but is not minimal. □

Horn theories have a well-known semantic characterization. A *model* is a vector $v \in \{0,1\}^n$, whose i-th component is denoted by v_i. For $B \subseteq \{1,\ldots,n\}$, we let x^B be the model v such that $v_i = 1$, if $i \in B$ and $v_i = 0$, if $i \notin B$, for $i \in \{1,\ldots,n\}$. The set of models of formula φ (resp. theory Σ), denoted by $mod(\varphi)$ (resp. $mod(\Sigma)$), is defined as usual.

For models v, w, we denote by $v \leq w$ the usual componentwise ordering, i.e., $v_i \leq w_i$ for all $i = 1,2,\ldots,n$, where $0 \leq 1$; $v < w$ means $v \neq w$ and $v \leq w$. For any set of models $\mathcal{M}$, we denote by $\max(\mathcal{M})$, (resp., $\min(\mathcal{M})$) the set of all maximal (resp., minimal) models in M. Denote by $v \wedge w$ componentwise AND of vectors $v, w \in \{0,1\}^n$, and by $Cl_\wedge(S)$ the closure of $S \subseteq \{0,1\}^n$ under $\wedge$. Then, a theory Σ is Horn representable, iff $mod(\Sigma) = Cl_\wedge(mod(\Sigma))$.

Example 2 Consider $\mathcal{M}_1 = \{(0101), (1001), (1000)\}$ and $\mathcal{M}_2 = \{(0101), (1001), (1000), (0001), (0000)\}$. Then, for $v = (0101)$, $w = (1000)$, we have $w, v \in \mathcal{M}_1$, while $v \wedge w = (0000) \notin \mathcal{M}_1$; hence $\mathcal{M}_1$ is not the set of models of a Horn theory. On the other hand, $Cl_\wedge(\mathcal{M}_2) = \mathcal{M}_2$, thus $\mathcal{M}_2 = mod(\Sigma)$ for some Horn theory Σ.

As discussed by Kautz *et al.* [12], a Horn theory Σ is semantically represented by its characteristic models, where $v \in mod(\Sigma)$ is called *characteristic* (or *extreme* [5]), if $v \notin Cl_\wedge(mod(\Sigma) \setminus \{v\})$. The set of all such models, the *characteristic set of* Σ, is denoted by $char(\Sigma)$. Note that $char(\Sigma)$ is unique. E.g., $(0101) \in char(\Sigma_2)$, while $(0000) \notin char(\Sigma_2)$; we have $char(\Sigma_2) = \mathcal{M}_1$.

3 Generating Explanations

In this section, we consider the generation of all explanations for an atom q. We exclude in our considerations the trivial explanation $E = \{q\}$, which always exists if q belongs to the assumption literals A, $\Sigma \cup \{q\}$ is satisfiable and $\Sigma \not\models q$. These conditions can be efficiently checked under both formula- and model-based representations.

Recall that a prime implicate (res., prime implicant) of a theory Σ is a smallest (w.r.t. inclusion) clause c (resp., term t) such that $\Sigma \models c$ (resp., $t \models \Sigma$). As well-known, explanations can be characterized by prime implicates as follows (cf., e.g., [18, 11]).

Proposition 2 *For a theory* Σ, E *is a nontrivial explanation of* q *w.r.t.* $A \subseteq Lit$ *if and only if the clause* $c = \bigvee_{p \in E} \overline{p} \vee q$ *is a prime implicate of* Σ *such that* $E \subseteq A$.

We start with the generation of all nontrivial explanations under formula-based representation. For this problem, we present the following algorithm.

Algorithm EXPLANATIONS

Input: A Horn CNF φ and a positive letter q.

Output: All nontrivial explanations of q from φ.

Step 1. $\varphi^\star := \emptyset$, $S := \emptyset$, and $O := \emptyset$;

Step 2. **for each** $c \in \varphi$ **do**
add any prime implicate $c' \subseteq c$ of φ to $\varphi^\star$;
for each $c' \in \varphi^\star$ with $P(c') = \{q\}$ and $N(c') \notin S$ **do**
begin output $N(c')$; $S := S \cup \{N(c')\}$;
$O := O \cup \{(c, c') \mid c \in \varphi^\star\}$
end;

Step 3. **while** some $(c_1, c_2) \in O$ exists **do**
begin $O := O \setminus \{(c_1, c_2)\}$;
if (1) $q \notin N(c_1)$, (2) $P(c_1) = \{r\} \subseteq N(c_2)$ and
(3) $\varphi^\star \cup N(c_1) \cup N(c_2) \setminus P(c_1)$ is satisfiable
then begin c := resolvent of c_1 and c_2;
compute any prime implicate $c' \subseteq c$ of φ;
if $N(c') \notin S$ **then**
begin output $N(c')$; $S := S \cup \{N(c')\}$;
$O := O \cup \{(c, c') \mid c \in \varphi^\star\}$
end
end
end. □

Example 3 We consider algorithm EXPLANATION on input of $\varphi = (\overline{x}_1 \vee \overline{x}_4)(\overline{x}_4 \vee \overline{x}_3)(\overline{x}_1 \vee x_2)(\overline{x}_3 \vee \overline{x}_5 \vee x_1)$, and $q = x_2$. As easily seen, each clause in φ is a prime implicate, and thus after Step 2, $\varphi^\star = \varphi$ and $S = \{\{x_1\}\}$. Furthermore, the explanation $E_1 = \{x_1\}$ was output.

In Step 3, (c_1, c_2), where $c_1 = \overline{x}_3 \vee \overline{x}_5 \vee x_1$ and $c_2 = \overline{x}_1 \vee x_2$, is the only pair in O satisfying (2) $P(c_2) = \{x_1\} \subseteq N(c_2)(= \{x_1\})$; moreover, (1) $q \notin N(c_1)(= \{x_3, x_5\})$ holds and (3) $\varphi^\star \cup \{x_3, x_5\}$ is satisfiable. Thus, in the body of the while-loop, its resolvent $c = \overline{x}_3 \vee \overline{x}_5 \vee x_2$ is computed. Clause c is a prime implicate of φ, and thus $E_2 = \{x_3, x_5\}$ is output and added to S. Furthermore, O is updated.

In the next iterations, no pair $(c_1, c_2) \in O$ is found which satisfies condition (2), and thus the algorithm halts. Note that E_1 and E_2 are the nontrivial explanations of $q = x_2$. □

The following result states that our algorithm works as desired. For any formula φ, denote by $||\varphi||$ its length, i.e., the number of literal occurrences in it.

Theorem 1 *Algorithm* EXPLANATIONS *incrementally outputs, without duplicates, all nontrivial explanations of q from φ. Moreover, the next output (respectively termination) occurs within $O(e \cdot m \cdot n \cdot ||\varphi||)$ time, where m is the number of clauses in φ, n the number of atoms, and e the number of explanations output so far.*

Proof. Only pairs (c, c') are added to O such that c' is a prime implicate of φ. Furthermore, by condition (3) in Step 3, each such c' must have $P(c') = \{q\}$. Thus, by Props. 1 and 2, algorithm EXPLANATIONS outputs only nontrivial (clearly different) explanations $E_1, E_2, \ldots, E_k$ for q.

To show that it outputs all nontrivial explanations E for q, assume that such an E is not output, i.e., $E \neq E_i, i = 1, \ldots, k$. Let φ_N be the CNF of all negative prime implicates of φ, and let

$\varphi' = \varphi_N \cup \{\bigvee_{p\in E_i} \overline{p} \vee q \mid i \in \{1,\ldots,k\}\}$. Since $c = \bigvee_{p\in E} \overline{p} \vee q$ is a prime implicate of φ and $c \notin \varphi'$, there exists a model $v \in mod(\varphi')$ such that $c(v) = 0$. Let w be a maximal such model, i.e., no model $u\,(> w)$ exists such that $u \in mod(\varphi')$ and $c(u) = 0$. Since $c(w) = 0$ implies $w \notin mod(\varphi)$, there exists a prime implicate c_1 in $\varphi^\star$ (when Step 2 is finished) such that $c_1(w) = 0$. Clearly $c_1 \notin \varphi'$, i.e., c_1 is of form $c_1 = \bigvee_{p\in N(c_1)} \overline{p} \vee x_i$ such that $x_i \neq q$. Moreover, we have $q \notin N(c_1)$ by $c(w) = c_1(w) = 0$. Consider now the model w' defined by $w'_i = 1$ and $w'_j = w_j$, for all $j \neq i$. Note that $c(w') = 0$, and by the maximality of w, there is a prime implicate $c_2 \in \varphi'$ such that $c_2(w') = 0$. Since $c_2(w) = 1$ and $c_2(w') = 0$, we have $x_i \in N(c_2)$. Since $q \notin N(c_1)$, a resolvent $c^\star$ of c_1 and c_2 thus exists.

We claim that $c^\star$ creates a new prime implicate $c' = \bigvee_{p\in N(c')} \overline{p} \vee q \subseteq c^\star$ of φ in Step 3, i.e., $N(c') \neq E_i, i = 1,\ldots,k$. To show this, note that $c^\star(w) = 0$ holds from $c_1(w) = c_2(w') = 0$. Thus, c_2 is not negative, which implies $c_2 = \bigvee_{p\in E_i} \overline{p} \vee q$ for some i. Indeed, otherwise $c^\star$ is negative, hence $c'(w) = 0$ for some $c' \in \varphi_N$, which contradicts the definition of w. This means that (c_1, c_2) has been added to O, and thus the resolvent $c^\star$ must be considered in Step 3. Moreover, $c^\star(w) = 0$ and $(\bigvee_{p\in E_i} \overline{p} \vee q)(w) = 1$ for all $i = 1,\ldots,k$ implies that $c^\star$ constructs an explanation $N(c') \neq E_i, i = 1,\ldots,k$, which proves our claim. However, this contradicts our assumption that the output is $E_1,\ldots,E_k$.

Thus, EXPLANATION is correct, and it remains to verify the time bound. Computing a prime implicate $c' \subseteq c$ of φ in Steps 2 and 3 is feasible in time $O(n \cdot \|\varphi\|)$, and thus the outputs in Step 2 occur with $O(m\cdot n\cdot \|\varphi\|)$ delay. As for Step 3, note the O contains only pairs (c_1, c_2) where $c_1 \in \varphi^\star$ and $c_2 = N(c_2) \cup \{q\}$ such that $N(c_2)$ was output, and each such pairs is added to O only once. Thus, the next output or termination follows within $e \cdot m$ runs of the while-loop, where e is the number of solutions output so far. The body of the loop can be done, using proper data structures, in $O(n \cdot \|\varphi\|)$ time (for checking $N(c_1) \notin S$ efficiently, we may store S in a prefix tree). Thus, the time until the next output resp. termination is bounded by $O(e \cdot m \cdot n \cdot \|\varphi\|)$. □

From this result, we obtain the following corollary.

Corollary 1 *Given a Horn CNF φ and a query q, computing $O(n^k)$ many explanations of q, where k is a constant, is possible in polynomial time.*

This corollary implies that the conjecture of Selman and Levesque [20, p. 266] that generating $O(n)$ many explanations of q is NP-hard, where n is the number of propositional letters in the language, is not true (unless P=NP). Note, however, that by the results of [20], computing $O(n)$ many or all assumption-based explanations from a Horn Σ is not possible in total polynomial time unless $\mathrm{P} = \mathrm{NP}$.

Let us now consider computing all explanations in the model-based setting.

Theorem 2 *Given the characteristic set $\mathcal{M} = char(\Sigma) \subseteq \{0,1\}^n$ of a Horn theory Σ, a query q, and $A \subseteq Lit$, computing the set of all explanations for q from Σ w.r.t. A is polynomial-time equivalent to dualizing a positive CNF.*

Here, polynomial-time equivalence means mutual polynomial-time transformability between deterministic functions, i.e., A reduces to B, if there a polynomial functions f, g s.t. for any input

I of A, $f(I)$ is an input of B, and if O is the output for $f(I)$, then $g(O)$ is the output of I, cf. [16]; we also request that O has size polynomial in the size of the output for I (if not, trivial reductions may exist). In our reduction, explanations correspond to clauses of the dual prime CNF and vice versa.

Proof. By Propositions 1 and 2, we need only compute all nontrivial explanations $E \subseteq A$ corresponding to prime implicates c of Σ s.t. $P(c) = \{q\}$ and $N(c) = E \subseteq A \cap P$.

We describe how the problem can be transformed to dualization of polynomially many positive CNFs $\varphi_1,\ldots,\varphi_k$, such that the clauses of the dual prime CNFs ψ_i for φ_i correspond to the explanations of q w.r.t. $A \cap P$ (equivalently, w.r.t. A). Thus, the problem is polynomially reducible to dualizing (in parallel) several positive CNFs φ_i. By simple methods, we can combine $\varphi_1,\ldots,\varphi_k$ into a single CNF φ (using further variables) such that the clauses of the dual prime CNF for φ correspond to the explanations of q w.r.t. A. (This step is of less interest in practice, since dualization of the individual φ_i is at the core of the computation.)

To construct the φ_i, we proceed as follows. Let $q = x_j$.

(1) Define $\mathcal{M}_i = \{v \in \mathcal{M} \mid v_j = i\}$, $i \in \{0, 1\}$.

(2) For every model $v \in \max(\mathcal{M}_1)$, let

$$F_v = \max(\{v \wedge x^A \wedge w \mid w \in \max(\mathcal{M}_0)\}).$$

We associate with F_v a monotone Boolean function f_v on the variables $P_v = A \cap \{x_i \mid v_i = 1\}$ such that $f_v(w) = 0 \Leftrightarrow w \leq s$ for some vector s in the projection of F_v on P_v. That is, F_v describes the maximal false points of f_v.

(3) Finally, define for every $v \in \max(\mathcal{M}_1)$

$$\varphi_v = \{c \mid N(c) = \emptyset,\, P(c) = P_v \setminus S,\, x^S \in F_v\}.$$

Note that φ_v is a prime CNF for f_v.

As shown in the Appendix, the nontrivial explanations of q w.r.t. A are given by the clauses in all dual prime CNFs ψ_v for φ_v where $v \in \max(\mathcal{M}_1)$ (equivalently, by all prime implicants t, i.e., a prime DNF representation of f_v, $v \in \max(\mathcal{M}_1)$). This proves one direction of the result.

For the converse, we show that, given a positive CNF φ on atoms P, computing an equivalent prime DNF ψ is reducible to computing all explanations as follows. Let q be a fresh letter (for component $n+1$), and define $\mathcal{M} = \{(v, 0) \mid v \in \max(\{w \mid \varphi(w) = 0\})\} \cup \{(11\cdots1)\}$ and $A = P$; note that $\max(\{w \mid \varphi(w) = 0\})$ is easily computed from φ. □

Example 4 Let $\mathcal{M} = \{(11011), (11010), (10101), (01010), (00001)\}$, and suppose we want all explanations of $q = x_1$ w.r.t. $A = \{x_3, x_4, x_5\}$. According to the above construction, we obtain:

(1) $\mathcal{M}_0 = \{(01010), (00001)\}$ and $\mathcal{M}_1 = \{(11011), (11010), (10101)\}$, thus we have $\max(\mathcal{M}_1) = \{(11011), (10101)\}$.

(2) We have two vectors $v^{(1)} = (11011)$ and $v^{(2)} = (10101)$:

$$
\begin{aligned}
F_{v^{(1)}} &= \max\left(\begin{Bmatrix}(11011)\wedge(00111)\wedge(01010),\\(11011)\wedge(00111)\wedge(00001)\end{Bmatrix}\right) = \{(00010),(00001)\},\\
F_{v^{(2)}} &= \max\left(\begin{Bmatrix}(10101)\wedge(00111)\wedge(01010),\\(10101)\wedge(00111)\wedge(00001)\end{Bmatrix}\right) = \{(00001)\}.
\end{aligned}
$$

Thus, $P_{v^{(1)}}=\{x_4, x_5\}$ and $f_{v^{(1)}}(w) = 0$ iff $w \in \{(10), (01), (00)\}$, and $P_{v^{(2)}}=\{x_3, x_5\}$ and $f_{v^{(2)}}(w) = 0$ iff $w \in \{(01), (00)\}$.

(3) We obtain $\varphi^{(1)} = x_4 \wedge x_5$ and $\varphi^{(2)} = x_3$. The respective prime dual CNFs are $\psi^{(1)} = x_4 \vee x_5$ and $\psi^{(2)} = x_3$.

Thus, the explanations of q w.r.t. A are $E_1 = \{x_4, x_5\}$ and $E_2 = \{x_3\}$. It can be seen that this is the correct result.

4 Negative Queries

So far, we considered Horn theories Σ and queries given by a letter q. In a general setting, we might allow that the formulas in Σ and the query are arbitrary propositional formulas. As for computation, we can introduce for a query, given by any formula χ, a fresh letter q, add implications $q \to \chi$, $\chi \to q$ in Σ, and then ask for a nontrivial explanation of q. Thus, positive letter queries do not constrain the expressivity of the framework. However, this technique does not work for Horn theories, if one of the implications $q \to \chi$, $\chi \to q$ is not Horn. In the simplest case, χ is a negative literal $\overline{q}$.

The next result tells us that already in this case, abduction from a Horn CNF is intractable. Recall that a Horn CNF φ is *acyclic*, if the graph on P with arcs from $x_i \in N(c)$ to $x_i \in P(c)$, $c \in \varphi$, has no directed cycle. We use the following easy lemma.

Lemma 1 *Given a Horn theory Σ and a conjunction $\alpha = \gamma_1 \wedge \cdots \wedge \gamma_l$ of clauses γ_i, deciding whether $\Sigma \models \alpha$ is possible in polynomial time.*

Proof. We have $\Sigma \models \alpha$ iff $\Sigma \models \gamma_i$, for each clause γ_i of α. Since clause inference from a Horn theory is polynomial, each test $\Sigma \models \gamma_i$ is polynomial and thus deciding $\Sigma \models \alpha$ us polynomial as well. □

Theorem 3 *Given a Horn CNF φ, a query χ in CNF, and $A \subseteq Lit$, deciding if χ has a nontrivial explanation w.r.t. A is* NP*-complete. Hardness holds even if $\chi = \overline{q}$, φ is acyclic, and either (i) $A = P$ or (ii) $A = P' \cup \overline{P}'$ for some $P' \subseteq P$.*

Proof. The problem is in NP, since an explanation exists iff some set $E \subseteq A$ exists such that the following two conditions hold: (i) $\Sigma \cup E$ is satisfiable and (ii) $\Sigma \cup E \models \chi$. A suitable E can be guessed and the conditions (i) and (ii) can be checked, by Lemma 1, in polynomial time.

Hardness is shown by a reduction from 3SAT. Let $\gamma = c_1 \wedge \cdots \wedge c_m$ be a 3CNF over atoms $x_1, \ldots, x_n$, where $c_i = \ell_{i,1} \vee \ell_{i,2} \vee \ell_{i,3}$. We introduce for each clause c_i a new atom y_i, for each x_j a new atom x'_j, and a special atom z. The Horn CNF φ contains the following clauses:

- $\overline{x}_i \vee \overline{x'_i}$, for all $i = 1, \ldots, n$;
- $\overline{z} \vee y_1$
- $\overline{y}_i \vee \overline{\ell}_{i,j} \vee y_{i+1}$ if $\ell_{i,j}$ is positive and $\overline{y}_i \vee \overline{\ell'_{i,j}} \vee y_{i+1}$ if $\ell_{i,j}$ is negative, for all $i = 1, \ldots, m-1$ and $j = 1, 2, 3$;
- $\overline{y}_m \vee \overline{\ell}_{m,j}$ if $\ell_{m,j}$ is positive and $\overline{y}_m \vee \overline{\ell'_{m,j}}$ if $\ell_{i,j}$ is negative, for $j = 1, 2, 3$.

As easily seen, φ is acyclic Horn. We claim that the query $q = \neg z$ has a nontrivial explanation E consisting of positive literals if and only if γ is satisfiable.

Let E be such an explanation. Then it follows from Proposition 2 that $\bigvee_{p \in E} \overline{p} \vee \overline{z}$ is a prime implicate of φ. Since all prime implicates can be generated by the resolution procedure, $E \subseteq \{x_1, \ldots, x_n, x'_1, \ldots, x'_n\}$ and E contains at least one atom $\ell^\star_{i,j}$ for every i, where $\ell^\star_{i,j} = \ell_{i,j}$ if $\ell_{i,j}$ is positive; otherwise, $\ell'_{i,j}$. Moreover, since φ contains clauses $\overline{x}_i \vee \overline{x'_i}$, $i = 1, \ldots, n$, E contains at most one atom in $\{x_i, x'_i\}$ for every i. Therefore, E corresponds to a satisfiable assignment for γ.

On the other hand, let x be a satisfiable assignment for γ. Then we arbitrarily choose from each c_i a literal ℓ_{i,j_i} satisfied by x. Let $\ell^\star_{i,j_i} = \ell_{i,j_i}$ if ℓ_{i,j_i} is positive; otherwise, ℓ'_{i,j_i}. Then $\bigvee_i \overline{\ell^\star_{i,j_i}} \vee \overline{z}$ is an implicant of φ. Note that $\bigvee_i \overline{\ell^\star_{i,j_i}}$ is not an implicant of φ, since it cannot be generated by the resolution procedure for φ. Thus we have a negative prime implicant containing $\overline{z}$. From Proposition 2, this shows the if-part. This proves NP-hardness under restriction (i).

For the restriction (ii), we use the same Horn CNF φ. Then, it is not hard to see that for $\chi = \overline{z}$ some explanation w.r.t. $A = \{x_i, x'_i, \overline{x}_i, \overline{x'_i} \mid 1 \leq i \leq n\}$ exists if and only if γ is satisfiable. $\Box$

Note that this result contrasts the tractability result that a nontrivial explanation $E \subseteq P$ for a positive query q can be computed in polynomial time [20]. Thus, the framework of Horn abduction is sensitive with respect to query representation. We also remark that we can find an arbitrary explanation E for a query $\overline{q}$ (which may contain negative literals), in polynomial time.

In the model-based setting, we obtain for computing all explanations for a negative query a similar result as for a positive query.

Theorem 4 *Given the characteristic set $\mathcal{M} = char(\Sigma) \subseteq \{0,1\}^n$ of a Horn theory Σ, a negative query $\overline{q}$, and $A \subseteq Lit$, computing all explanations for q from Σ w.r.t. A is polynomial-time equivalent to dualizing a positive CNF.*

Proof. Observe that since the query is not positive, explanations of q may involve negative literals.

Proposition 2 implies that the nontrivial explanations for $\overline{q}$ w.r.t. A correspond to the prime implicates c of Σ such that $q \in N(c)$ and $P(c) \cup N(c) \subseteq A \cup \{q\}$. Let $q = x_j$, and define sets $\mathcal{M}_0$ and $\mathcal{M}_1$ for $\mathcal{M}$ as in the proof of Theorem 2. Denote by A_+ (resp., A_-) the set of positive (resp., negative) literals in A. We consider the following two cases:

(1) *Positive explanations for* $\overline{q}$. I.e., all prime implicates c of Σ s.t. $\{q\} \subseteq N(c) \subseteq A_+ \cup \{q\}$ and $P(c) = \emptyset$.

Similarly as in the proof of Theorem 2, we construct dualization problems for functions f_v, but for $v \in \max(\mathcal{M}_0)$:

(1.1) For every $v \in \max(\mathcal{M}_0)$, let

$$F_v = \max(\{v \wedge x^{A_+} \wedge w \mid w \in \max(\mathcal{M}_1)\}).$$

The associated monotone Boolean function f_v on $P_v = A \cap \{x_i \mid v_i = 1\}$ is defined by $f_v(w) = 0 \Leftrightarrow w \leq s$ holds for some vector s in the projection of F_v on P_v.

(1.2) We define, for $v \in \max(\mathcal{M}_0)$,

$$\varphi_v = \{c \mid N(c) = \emptyset,\, P(c) = P_v \setminus S,\, x^S \in F_v\}.$$

Similarly as in Theorem 2, we can show that the clauses in the dual prime CNFs for all φ_v, $v \in \max(\mathcal{M}_0)$, correspond to the positive explanations of $\overline{q}$ (cf. Appendix).

(2) *Non-positive explanations for* $\overline{q}$. These are all prime implicates c of Σ s.t. $\{q\} \subseteq N(c) \subseteq A_+ \cup \{q\}$ and $P(c) = \{r\}$, where $r \in A_-$.

For each $r = x_{j'}$ (where $j' \neq j$), we proceed as follows.

(2.1) For every $v \in \max(\mathcal{M}_0)$ and $i \in \{0, 1\}$, define

$$\mathcal{M}_i^r = \{v \in \mathcal{M}_i \mid v_r = i\}.$$

(2.2) For each $v^{(0)} \in \max(\mathcal{M}_0^r)$ and $v^{(1)} \in \max(\mathcal{M}_1^r)$, let

$$F_{v^{(0)},v^{(1)}} = \max(\{v^{(0)} \wedge v^{(1)} \wedge x^{A_+} \wedge w \mid w \in \max(\{u \in \mathcal{M}_1 \mid u_{j'} = 0\})\}).$$

We associate with it a monotone Boolean function $f_{v^{(0)},v^{(1)}}$ on $P_{v^{(0)},v^{(1)}} = A \cap \{x_i \mid v_i^{(0)} = v_i^{(1)} = 1\}$ such that $f_{v^{(0)},v^{(1)}}(w) = 0 \Leftrightarrow w \leq s$ for some s in the projection of $F_{v^{(0)},v^{(1)}}$ on $P_{v^{(0)},v^{(1)}}$.

(2.3) We define the CNFs

$$\varphi_{v^{(0)},v^{(1)}} = \{c \mid N(c) = \emptyset,\, P(c) = P_v \setminus S,\, x^S \in F_{v^{(0)},v^{(1)}}\}.$$

As shown in the Appendix, the clauses in the dual prime CNFs $\psi_{v^{(0)},v^{(1)}}$ for all $\varphi_{v^{(0)},v^{(1)}}$, where $v^{(0)} \in \max(\mathcal{M}_0^r)$ and $v^{(1)} \in \max(\mathcal{M}_1^r)$ and $r \in A_-$, correspond to the non-positive explanations of $\overline{q}$.

In total, computing all explanations of $\overline{q}$ is polynomial-time reducible to dualizing (in parallel) polynomially many positive CNFs. As mentioned in the proof of Theorem 2, this is polynomially reducible to dualizing a single CNF.

The converse is shown by a reduction similar to the one in the proof of Theorem 2; we just invert the polarity of q (i.e., interchange 0 and 1 in the position of q). $\square$

5 Joint Explanations

We call any set of $E \subseteq A$ of literals a *joint explanation* of observations $o_1, o_2, \ldots, o_l$ from a background theory Σ w.r.t. a set of assumptions $A \subseteq Lit$, if E is an explanation of each o_i from Σ w.r.t. A. The observations o_i may be letters, or in a generalized setting propositional formulas.

Note that any such E is also an explanation for the conjunction $\alpha = o_1 \wedge o_2 \wedge \cdots \wedge o_l$ of all observations, while the converse is not true in general: an explanation E of α may not satisfy minimality for o_1, say, i.e., some $E' \subset E$ may explain o_1. Thus, joint explanations are stronger than ordinary explanations. In case of multiple explanations, this may be used to single out those which match with each of the (possibly independently made) observations.

For example, the malfunctioning of a car may be explained by two car mechanics, based on observations o_1 and o_2, respectively having the same background knowledge. A match of their (individual) diagnoses E_1 and E_2 (i.e., $E_1 = E_2$) might be taken in favor of believing their correctness. In fact, each diagnosis E_i is robust in the sense that adding the other observation does not require a change of the diagnosis; from another perspective, the same diagnosis is good for explaning different observations. If E_1 and E_2 are different, then we might want to know whether alternative diagnoses E_1' and E_2' do exist which coincide, i.e., whether a joint explanation is possible.

We now present some results on the complexity of joint explanations. In that, we use the following characterisation of explanations, which is immediate by monotonicity of classical logic.

Lemma 2 *For any theory Σ, query χ, and any $E, A \subseteq Lit$, E is an explanation for χ from Σ w.r.t. A iff the following conditions hold: (i) $\Sigma \cup E$ is satisfiable, (ii) $\Sigma \cup E \models \chi$, and (iii) $\Sigma \cup (E \setminus \{\ell\}) \not\models \chi$, for every $\ell \in E$.*

Thus, by combining Lemmas 1 and 2, we obtain the following result on recognizing joint explanations.

Proposition 3 *Given a Horn theory Σ, a query χ in CNF, and sets $E, A \subseteq Lit$, deciding whether E is an explanation for χ from Σ w.r.t. A is possible in polynomial time.*

Thus, the recognition problem is not harder than for ordinary explanations. However, deciding the existence of some joint explanation is harder.

Theorem 5 *Given a Horn CNF φ, query CNFs $\chi_1, \chi_2, \ldots, \chi_l$, where $l \geq 2$, and$A \subseteq Lit$, deciding if a joint explanation exists from Σ w.r.t. A is* NP*-complete. Hardness holds even if $l = 2$, each χ_i is a letter, φ is acyclic, and $A = Lit$.*

Proof. The problem is in NP, since we can guess $E \subseteq A$ and check in polynomial time whether E is an explanation for each χ_i w.r.t. A, $i \in \{1, \ldots, l\}$ by Proposition 3.

The hardness part is shown by a reduction from 3SAT. Let $\gamma = c_1 \wedge \cdots \wedge c_m$ be a 3CNF on $x_1, \ldots, x_n$. We take fresh atoms $x_{1'}$, $z_1, \ldots, x_{n'}$, $z_{n'}$, $x_{c_1}, \ldots, x_{c_{m+n}}$, x_{n+1}, x_{n+2} and define a Horn CNF φ with the following clauses:

- $\bigvee_i \overline{z}_i \vee x_{n+1}$;

- $\overline{x}_i \vee z_i$ and $\overline{x}_{i'} \vee z_i$, $i = 1, 2, \ldots, n$;
- $\bigvee_j \overline{x}_{c_j} \vee x_{n+2}$;
- $\overline{x}_i \vee x_{c_j}$, if $x_i \in c_j$, and $\overline{x}_{i'} \vee x_{c_j}$ if $\overline{x}_i \in c_j$, $j = 1, \ldots, m$.
- $\overline{x}_i \vee x_{c_{m+i}}$, $\overline{x}_{i'} \vee c_{m+i}$, $i = 1, \ldots, n$.

As easily seen, φ is acyclic Horn. We claim that a joint explanation for $q_1 = x_{n+1}$ and $q_2 = x_{n+2}$ exists if and only if γ is satisfiable.

Let E be a joint explanation for x_{n+1} and x_{n+2}. Then since E is an explanation for x_{n+1}, we have $|E \cap \{x_i, x_{i'}, z_i\}| = 1$ for every $i = 1, \ldots, n$. Since E is an explanation for x_{n+2}, we have $|E \cap (\{x_i \mid x_i \in c_j\} \cup \{x_{i'} \mid \overline{x}_i \in c_j\} \cup \{x_{c_j}\})| \geq 1$ for every $j = 1, \ldots, m$, and $|E \cap \{x_i, x_{i'}, x_{c_{m+i}}\}| \geq 1$ for every $i = 1, \ldots, n$. Thus $|E \cap \{x_i, x_{i'}\}| = 1$ for $i = 1, \ldots, n$ and $|E \cap (\{x_i \mid x_i \in c_j\} \cup \{x_{i'} \mid \overline{x}_i \in c_j\})| \geq 1$ for every $j = 1, \ldots, m$. This means that E corresponds to a satisfiable assignment for γ.

On the other hand, let y be a satisfiable assignment for φ. Then define $E = \{x_i \mid y_i = 1, i = 1, \ldots, n\} \cup \{x_{i'} \mid y_i = 0, i = 1, \ldots, n\}$. We can easily see that this E is a joint explanation for x_{n+1} and x_{n+2}. □

For model-based representation, we obtain similar results.

Proposition 4 *Given the characteristic set $\mathcal{M} = char(\Sigma)$ of a Horn theory Σ, a query χ in CNF, and sets $E, A \subseteq Lit$, deciding whether E is an explanation for χ from Σ w.r.t. A is possible in polynomial time.*

Proof. By Lemma 2, E is an explanation for χ iff the following conditions hold: (i) $\Sigma \cup E$ is satisfiable, (ii) $\Sigma \cup E \models \chi$, and (iii) $\Sigma \cup (E \setminus \{\ell\}) \not\models \chi$, for every $\ell \in E$.

Each of these tests is feasible in polynomial time, which can be derived from the result that given the characteristic set $char(\Sigma_i)$ of Horn theories Σ_i, $i \in \{1, 2\}$ and a CNF χ, deciding whether $\Sigma_1 \cup \Sigma_2 \models \chi$ is possible in polynomial time [9, Corollary 6.3], and the fact that any set E of literals is Horn and its characteristic set is easily computed: if E is consistent, then

$$char(E) = \left\{x^{\{1,\ldots,n\}\setminus\{i \mid \overline{x}_i \in E\}}\right\} \cup \left\{x^{\{1,\ldots,n\}\setminus(\{i \mid \overline{x}_i \in E\} \cup \{j\})} \mid x_j, \overline{x}_j \notin E\right\};$$

otherwise, $char(E) = \emptyset$. □

Theorem 6 *Given the characteristic set $\mathcal{M} = char(\Sigma)$ of a Horn theory Σ, query CNFs $\chi_1, \chi_2, \ldots, \chi_l$, $l \geq 2$, and $A \subseteq Lit$, deciding if a joint explanation exists from Σ w.r.t. A is* NP*-complete. Hardness holds even if $l = 2$, each χ_i is a letter, and $A = Lit$.*

Proof. The problem is in NP, since we can guess a joint explanation E for $\chi_1, \ldots, \chi_l$ and check, by Proposition 4, in polynomial time whether E is an explanation for χ_i w.r.t. A, for each $i = 1, \ldots, l$.

NP-hardness for the stated restriction is shown by a reduction from 3SAT. Let $\varphi = c_1 \wedge \cdots \wedge c_m$ be a 3CNF on $x_1, \ldots, x_n$. We take fresh atoms $x_{1'}$,…,$x_{n'}$, x_{n+1}, x_{n+2} and define the following sets of models:

$$\begin{aligned}
\mathcal{M}_1 &= \left\{x^{\{1,1',\ldots,n,n',n+2\}\setminus\{i,i'\}} \mid 1 \leq i \leq n\right\}, \\
\mathcal{M}_2 &= \left\{x^{\{1,1',\ldots,n,n',n+1\}\setminus\{i,k' \mid x_i \in c_j, \overline{x}_k \in c_j\}} \mid 1 \leq j \leq m\right\}, \\
\mathcal{M}_3 &= \left\{x^{\{1,1',\ldots,n,n',n+1\}\setminus\{i,i'\}} \mid 1 \leq i \leq n\right\}.
\end{aligned}$$

Let $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3 \cup \{(11\cdots1)\}$. Note that $\mathcal{M}$ is the characteristic set of some Horn theory Σ. Intuitively, $\mathcal{M}_1$ generates all truth assignments to $x_1, \ldots, x_n$ as explanations of x_{n+1}, and $M_2 \cup \mathcal{M}_3$ generates all minimal selections of at least one literal in each $c_j \in \varphi$, $j = 1, \ldots, m$, and each $\{x_i, \overline{x}_i\}$, $i = 1, \ldots, n$, as explanations of x_{n+2}. The model $(11\cdots1)$ ensures consistency of the explanation candidates.

We claim that a joint explanation E for $q_1 = x_{n+1}$ and $q_2 = x_{n+2}$ exists if and only if φ is satisfiable.

Let E be a joint explanation for x_{n+1} and x_{n+2}. Observe that any such E must consist, by Proposition 1, of positive literals only. Since E is an explanation for x_{n+1}, we have $x_{n+2} \notin E$ and $|E \cap \{x_i, x_{i'}\}| = 1$ for $i = 1, 2, \ldots n$. Moreover, since E is an explanation for x_{n+2}, we have $x_{n+1} \notin E$, $|E \cap \{x_i, x_{i'}\}| \geq 1$ for $i = 1, 2, \ldots n$, and $|E \cap \{x_i, x_{k'} \mid x_i \in c_j, \overline{x}_k \in c_j\}| \geq 1$ for $j = 1, 2, \ldots m$. Thus E corresponds to a satisfiable assignment for φ.

On the other hand, let y be a satisfiable assignment for φ. Then define $E = \{x_i \mid y_i = 1, i = 1, \ldots, n\} \cup \{x_{i'} \mid y_i = 0, i = 1, \ldots, n\}$. We can easily see that this E is a joint explanation for x_{n+1} and x_{n+2}. $\Box$

Thus, the tractability results in [20, 12] do not generalize to joint explanations for positive queries. Similar intractability results hold for negative queries and combined positive and negative queries.

6 Related Works

Selman and Levesque [19, 20] were among the first to study the complexity of computing general and assumption-based explanations; Corollary 1 closes an open problem of them. The underlying algorithm EXPLANATIONS is a relative of a similar procedure by Boros *et al.* [2] for computing all prime implicates of a Horn CNF in output-polynomial time. In fact, Theorem 1 can be seen as a strengthening of their result. For negative queries, a similar algorithm is not evident.

del Val [6] presented generation of implicates and prime implicates of certain clausal theories in a target language, which is formed on a subset of the atoms, using a procedure based on kernel resolution and derived exponential bounds on its running time. Furthermore, del Val presented formulas for bounding the number of prime implicates in a target language and described the use of this procedure for generating jointly all explanations of all literals which are not on a set of atoms V. However, neither is this method incremental in nature, nor is it clear whether it is total polynomial time. Furthermore, it considers a letter q and its negation $\overline{q}$ at once.

Inoue [11] considered, in the propositional and the first-order context, generating explanations and prime implicates using SOL-resolution. He proposed a strategy which processes, starting from the empty set, clauses from a theory incrementally. However, due to possible large intermediate results, this method is not total polynomial time in general.

Khardon *et al.* [13] show how computing all keys of a relational database schema, which is constrained by a Boolean formula φ, can be polynomially transformed into computing all explanations of a query q from $\varphi \wedge \psi$, where ψ is Horn. Thus, our algorithm EXPLANATIONS can be used for efficiently generating all keys of a database scheme where φ is a Horn CNF. This generalizes the result for φ consisting of non-negative Horn clauses, i.e., a set of functional dependencies [15].[1] Note that Khardon *et al.* also show how to compute a single explanation of a query q from a theory φ polynomially, using repeatedly an oracle for computing a key of a database schema constrained by $\varphi \wedge \psi$, where ψ is Horn; however, this method is not usable for generating explanations from general Horn theories (cf. Footnote 1).

Less related to our work is the one by Eiter and Gottlob [7], who considered abduction from general and Horn propositional theories. However, they focused on existence of explanations and reasoning tasks about explanations, and did not consider generating multiple explanations.

7 Conclusion

We have presented a number of positive and negative results about generating all and some abductive explanations, respectively, which complement previous work in the literature. In particular, we analyzed the role of positive vs negative abductive queries, under both formula- and model-based representation, and we considered the novel notion of joint explanation. Our positive results may be readily applied for efficiently computing (a subset of) all explanations. The results draw a complete picture for the model-based setting, and almost so for the formula-based setting; the complexity of generating all explanations for a negative query in it is currently open.

A Appendix

Proof of Theorem 2 (cont'd) Let $q = x_j$, let c be a prime implicate of Σ such that $P(c) = \{q\}$ and $N(c) \subseteq A \cap P$, and let $t = \bigwedge_{p \in N(c)} p$. Then since c is a prime implicate of Σ, there exists a model $v \in mod(\Sigma)$ such that $v_j = 1$ and $t(v) = 1$. Since any model $w \in mod(\Sigma)$ with $w \geq v$ also satisfies $w_j = 1$ and $t(w) = 1$, we can choose such models v from $\max(\mathcal{M}_1)$. Since $N(c) \subseteq A$, we have

$$t(v \wedge x^A) \quad = \quad 1 \quad \text{for some} \quad v \in \max(\mathcal{M}_1). \tag{1}$$

Moreover, no model $w \in mod(\Sigma)$ satisfies $w_j = 0$ and $t(w) = 1$. By definition of $F_v = \max(\{v \wedge x^A \wedge w \mid w \in \max(\mathcal{M}_0)\})$ in (2), we have

$$t(u) \quad = \quad 0 \quad \text{for all} \quad u \in F_v. \tag{2}$$

[1] In fact, Khardon *et al.*'s transformation works only if φ has no negative prime implicate; otherwise, keys introduce inconsistency.

On the other hand, for the associated monotone Boolean function f_v on $P_v = A \cap \{x_i \mid v_i = 1\}$, any prime implicant t (i.e., any minimal positive term t satisfying (1) and (2)) produces a prime implicate c of Σ s.t. $P(c) = \{q\}$ and $N(c) = P(t) \subseteq A \cap P$.

The prime implicants of f_v correspond to the clauses in the (unique) prime CNF ψ_v describing the dual function of f_v. As easily seen, the CNF φ_v correctly represents f_v.

Therefore, the problem is polynomial-time reducible to dualizing (in parallel) several positive CNFs φ_v, $v \in \max(\mathcal{M}_1)$.

The (parallel) dualization of arbitrary positive CNFs $\varphi_1, \ldots, \varphi_m$, $m \geq 2$, on atoms $P = \{x_1, \ldots, x_n\}$ can be polynomially reduced to dualization of a single positive CNF φ as follows. Without loss of generality, each φ_i is nonempty and does not contain the empty clause.

Let $t = x_{i_1} x_{i_2} \cdots x_{i_\ell}$ be any prime implicant of φ_1, and let $Y = \{y_1, \ldots, y_m\}$, $Z = \{z_{i_1}, \ldots, z_{i_\ell}\}$ be sets of fresh atoms. Define

$$\varphi \quad = \quad \bigwedge_{j=1}^{\ell} z_{i_j} \wedge \bigwedge_{i=1}^{m} \bigwedge_{c \in \varphi_i} (c \vee y_i) \wedge \bigwedge_{1 \leq i < j \leq m} (y_i \vee y_j)$$

Clearly, φ is constructible in polynomial time. Then, c is a clause in the dual prime CNF of φ_i iff $c \cup \{z_{i_1}, \ldots, z_{i_\ell}\} \cup Y \setminus \{y_i\}$ is a clause in the dual prime CNF of φ, for $i \in \{1, \ldots, m\}$. The only clause in the dual prime CNF of φ not of this form is $\{z_{i_1}, \ldots, z_{i_\ell}\} \cup Y$, which we let correspond to $x_{i_1} \vee \cdots \vee x_{i_\ell}$.

Note that by this definition, for each clause c in the dual prime CNF of φ, the corresponding clause from the dual prime CNFs of all φ_i is entirely encoded in c itself. The purpose of the z_{i_j} variables is to provide a meaningful encoding for the case where Y is included in c. Note that the size of the dual prime CNF of φ is polynomial in the size of the dual prime CNFs of all φ_i. Therefore, a polynomial time reduction using functions f, g as described is easily defined). □

Proof of Theorem 4 (cont'd) (1) Positive explanations for $\overline{q}$, i.e., negative prime implicates c of Σ s.t. $\{q\} \subseteq N(c) \subseteq A_+ \cup \{q\}$ and $P(c) = \emptyset$.

Let $t = \bigwedge_{p \in N(c) \setminus \{q\}} p$. Since c is a prime implicate of Σ, there exists a model $v \in mod(\Sigma)$ such that $v_j = 0$ and $t(v) = 1$, Since any model $w \in mod(\Sigma)$ with $w \geq v$ also satisfies $t(w) = 1$, we can choose such models v from $\max(\mathcal{M}_0)$, Since $N(c) \subseteq A_+$, we have

$$t(v \wedge x^{A_+}) \quad = \quad 1 \quad \text{for some } \ v \in \max(\mathcal{M}_0). \tag{3}$$

Moreover, no model $w \in mod(\Sigma)$ satisfies $w_j = 1$ and $t(w) = 1$. We have

$$t(u) \quad = \quad 0 \quad \text{for all } \ u \in F_v, \tag{4}$$

where F_v is defined as in (1.1). Any prime implicant t of the associated monotone Boolean function f_v (i.e., any minimal monotone term t satisfying (3) and (4)) produces a desired prime implicate c of Σ. Therefore, similarly as in the proof of Theorem 2, generating all positive explanations is polynomial-time reducible to the dualization problems for all φ_v as defined in (1.2), where $v \in \max(\mathcal{M}_0)$.

(2) Non-positive explanations for $\overline{q}$. Those are all prime implicates c of Σ s.t. $\{q\} \subseteq N(c) \subseteq A_+ \cup \{q\}$ and $P(c) = \{r\}$, where $r \in A_-$.

Let $q = x_j$ and $r = x_{j'}$, and let $t = \bigwedge_{p \in N(c) \setminus \{q\}} p$. Since c is a prime implicate of Σ, there exist two models $u^{(0)}, u^{(1)} \in mod(\Sigma)$ such that $u^{(0)}_j = 0$, $u^{(0)}_{j'} = 0$, $u^{(1)}_j = 1$, $u^{(1)}_{j'} = 1$, and $t(u^{(0)}) = t(u^{(1)}) = 1$. This means that

$$t(v^{(0)} \wedge v^{(1)} \wedge x^{A_+}) = 1, \quad \text{for some } v^{(0)} \in \max(\mathcal{M}^r_0) \text{ and } v^{(1)} \in \max(\mathcal{M}^r_1), \tag{5}$$

where $\max(\mathcal{M}^r_1)$ is defined as in (2.1). Moreover, no model $w \in mod(\Sigma)$ satisfies $w_j = 1$ $w_{j'} = 0$ and $t(w) = 1$. We have

$$t(u) \;=\; 0 \quad \text{for all } \; u \in F_{v^{(0)},v^{(1)}}, \tag{6}$$

where $F_{v^{(0)},v^{(1)}}$ is defined as in (2.2). On the other hand, any prime implicant t of the associated monotone Boolean function $f_{v^{(0)},v^{(1)}}$ (i.e., any minimal monotone term t satisfying (5) and (6)) produces a desired prime implicate c of Σ. Therefore, similarly as in the proof of Theorem 2, generating all non-positive explanations is polynomial-time reducible to the dualization problems for the CNFs $\varphi_{v^{(0)},v^{(1)}}$ in (2.3), $v^{(0)} \in \max(\mathcal{M}^r_0)$ and $v^{(1)} \in \max(\mathcal{M}^r_1)$. $\square$

References

[1] C. Bioch and T. Ibaraki. Complexity of identification and dualization of positive Boolean functions. *Information and Computation*, 123:50–63, 1995.

[2] E. Boros, Y. Crama, and P. L. Hammer. Polynomial-time inference of all valid implications for Horn and related formulae. *Annals of Mathematics and Artificial Intelligence*, 1:21–32, 1990.

[3] G. Brewka, J. Dix, and K. Konolige. *Nonmonotonic Reasoning – An Overview*. Number 73 in CSLI Lecture Notes. CSLI Publications, Stanford University, 1997.

[4] J. de Kleer. An assumption-based truth maintenance system. *Artificial Intelligence*, 28:127–162, 1986.

[5] R. Dechter and J. Pearl. Structure identification in relational data. *Artificial Intelligence*, 58:237–270, 1992.

[6] A. del Val. The complexity of restricted consequence finding and abduction. In *Proceedings AAAI-00*, pages 337–342, 2000.

[7] T. Eiter and G. Gottlob. The complexity of logic-based abduction. *Journal of the ACM*, 42(1):3–42, January 1995.

[8] T. Eiter, G. Gottlob, and K. Makino. New results on monotone dualization and generating hypergraph transversals. In *Proceedings 34th ACM Symposium on Theory of Computing (STOC 2002)*. ACM, 2002.

[9] T. Eiter, K. Makino, and T. Ibaraki. Computing intersections of Horn theories for reasoning with models. *Artificial Intelligence*, 110(1-2):57–101, 1999.

[10] M. Fredman and L. Khachiyan. On the complexity of dualization of monotone disjunctive normal forms. *Journal of Algorithms*, 21:618–628, 1996.

[11] K. Inoue. Linear resolution for consequence finding. *Artificial Intelligence*, 56(2-3):301–354, 1992.

[12] H. Kautz, M. Kearns, and B. Selman. Reasoning with characteristic models. In *Proceedings AAAI-93*, pages 34–39, 1993.

[13] R. Khardon, H. Mannila, and D. Roth. Reasoning with examples: Propositional formulae and database dependencies. *Acta Informatica*, 36(4):267–286, 1999.

[14] R. Khardon and D. Roth. Reasoning with models. *Artificial Intelligence*, 87(1/2):187–213, 1996.

[15] C. L. Lucchesi and S. Osborn. Candidate keys for relations. *Journal of Computer and System Sciences*, 17:270–279, 1978.

[16] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[17] D. Poole. Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence*, 5(1):97–110, 1989.

[18] R. Reiter and J. de Kleer. Foundations of assumption-based truth maintenance systems: Preliminary report. In *Proceedings AAAI-87*, pages 183–188, 1982.

[19] B. Selman and H. J. Levesque. Abductive and default reasoning: A computational core. In *Proceedings AAAI-90*, pages 343–348, July 1990.

[20] B. Selman and H. J. Levesque. Support set selection for abductive and default reasoning. *Artificial Intelligence*, 82:259–272, 1996.