Deriving Explanations for Decision Trees: The Impact of Domain Theories

Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, Nicolas Szczepanski

Deriving Explanations for Decision Trees: The Impact of Domain Theories

Gilles Audemard[1] , **Jean-Marie Lagniez**[1] , **Pierre Marquis**[1,2] , **Nicolas Szczepanski**[3]
[1]Univ. Artois, CNRS, CRIL, Lens, France
[2]Institut Universitaire de France
[3]IRT SystemX, Palaiseau, France

Abstract

We are interested in identifying the complexity of computing explanations of various types for a decision tree, when the Boolean conditions used in the tree are not independent. When a domain theory indicating how the Boolean conditions occurring in the tree are logically connected is available, taking advantage of it is important to derive provably correct explanations. In this paper, we show that leveraging such a domain theory may have a strong impact on the complexity of generating explanations. While computing a subset-minimal abductive explanation (or a subset-minimal contrastive explanation) for an instance becomes NP-hard in presence of a domain theory, tractable restrictions exist. Especially, domain theories expressing the encoding of numerical attributes into Boolean conditions lead to tractable explanation problems for contrastive explanations.

1 Introduction

Motivations Several types of explanations can be defined when dealing with AI systems that implement classifiers f, i.e., mappings from a set $\boldsymbol{X}$ of instances to a set $\mathcal{L}$ of classes. On the one hand, *abductive explanations* (see e.g., [Ignatiev *et al.*, 2019]) aim to explain the classification of an instance $\boldsymbol{x} \in \boldsymbol{X}$ as achieved by f. On the other hand, *contrastive explanations* (see e.g., [Miller, 2019]) aim to explain why an input instance $\boldsymbol{x}$ has *not* been classified by f as expected by the explainee. In both cases, explanations can be represented as subsets of the *characteristics* (i.e., the pairs attribute-value) of $\boldsymbol{x}$: in the case of abductive explanations, the subset of characteristics that is derived must be sufficient to justify the classification $f(\boldsymbol{x})$ that is made, in the sense that any instance sharing this subset of characteristics must be classified by f in the same way as $\boldsymbol{x}$; in the case of contrastive explanations, one is interested in pointing out a subset of characteristics of $\boldsymbol{x}$ that must be modified to get an instance $\boldsymbol{x}'$ satisfying $f(\boldsymbol{x}') \neq f(\boldsymbol{x})$.

In this paper, the problem of deriving *abductive / contrastive explanations suited to classifiers f (i.e., $\mathcal{L} = \{0, 1\}$) represented by decision trees* is considered. We focus on the decision tree model because it is considered as one of the leading forms of interpretable models, so that more opaque models can be distilled into decision trees to benefit from their improved interpretability [Ras *et al.*, 2022]. Indeed, a number of explanation and verification queries for decision trees can be answered using polynomial-time algorithms, while the same queries are intractable for many other ML models [Audemard *et al.*, 2021a].

When f is a decision tree [Breiman *et al.*, 1984; Quinlan, 1986], *decision nodes* over attributes A_i from $\mathcal{A}$ are used in the representation of f. Whenever A_i is numerical, the Boolean conditions labelling the nodes over A_i used in f take the form $(A_i \geq v_j^i)$. Whenever A_i is categorical and it has been one-hot encoded, the Boolean conditions labelling the nodes over A_i used in f take the form $(A_i = v_j^i)$.

A key observation is that *two spaces of characteristics* can be used to describe the instances and their explanations when f is a decision tree (and more generally, when f is a tree-based classifier, e.g., a random forest [Breiman, 2001], or a boosted tree [Freund and Schapire, 1997; Schapire and Freund, 2014; Friedman, 2001]). Indeed, instances and explanations can be represented as *sets of characteristics based on the initial set of attributes*, but also as *sets of characteristics based on the Boolean conditions used in f*. It turns out that considering the latter space of characteristics is preferable from an XAI perspective since it leads to explanations (abductive or contrastive) that are *more general* than those defined when the set of characteristics based on the initial set of attributes is considered (they cover more instances).

Let us illustrate it on a very simple loan granting scenario. Suppose that the decision tree classifier f, depicted on Figure 1, is used to determine whether the loan must be granted or not.

Alice wants to get a loan. Two attributes are used primarily to describe instances: A_1 (numerical) gives the annual incomes of the applicant, and A_2 (Boolean) indicates whether the applicant has reimbursed a previous loan. Alice's annual incomes are equal to $\$45\ k$ and she has reimbursed a previous loan. Thus, Alice corresponds to the instance $\boldsymbol{x} = (45, 1)$. Since $f(\boldsymbol{x}) = 1$, Alice will get the loan. The unique subset-minimal abductive explanation for $\boldsymbol{x}$ given f in the space of characteristics considered at start is $\{(A_1 = 45)\}$. Using words, the abductive explanation provided to Alice is "you got the loan since your annual incomes are equal to $\$45\ k$". In the space of characteristics of the predictor, two subset-

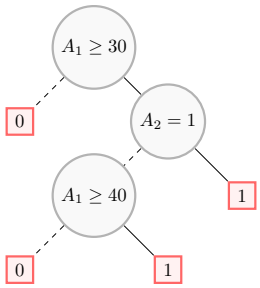

Figure 1: A simple decision tree classifier for granting loans.

minimal abductive explanations for $\boldsymbol{x}$ given f can be derived, namely $\{(A_1 \geq 40)\}$ and $\{(A_1 \geq 30), (A_2 = 1)\}$. Those explanations are better than the previous one $\{(A_1 = 45)\}$ since they correspond to more general classification rules and they reflect in a much more accurate way the behaviour of the predictor. Using words, "you got the loan since your annual incomes are greater than or equal to \$40 k, but also because your annual incomes are greater than or equal to \$30 k and you have reimbursed a previous loan".

Consider now Bob, who also wants to get a loan. Bob has reimbursed a previous loan, but his annual incomes are equal to \$20 k, only. Bob corresponds to the instance $\boldsymbol{x}' = (20, 1)$. Since $f(\boldsymbol{x}') = 0$, Bob will not get the loan. Using the definition provided in [Ignatiev *et al.*, 2020], the unique subset-minimal contrastive explanation for $\boldsymbol{x}'$ given f is $\{A_1\}$. Using words, "in order to get the loan, you have to change your annual incomes". This is correct, but clearly insufficient since Bob surely expects to know to which extent his annual incomes must be updated in order to get the loan. The contrastive explanation $\{\overline{(A \geq 30)}\}$, represented in the space of characteristics of the predictor, is a far better explanation. Indeed, it indicates that "in order to get the loan, you have to make your annual incomes at least equal to \$30 k". For more details on contrastive explanations for tree-based classifiers, see [Audemard *et al.*, 2023].

Contributions In the following, we look for explanations represented in the space of characteristics of the decision tree f to take advantage of their generality. Accordingly, f is now considered as a Boolean function based on the Boolean conditions labelling its decision nodes. However, f may be based on Boolean conditions that are *not logically independent*, especially when they come from the same (non-Boolean) attribute A_i used to describe instances at start. This is the case in the above example, where the Boolean conditions $(A_1 \geq 30)$ and $(A_1 \geq 40)$ are not independent, since no instance may satisfy $(A_1 \geq 40)$ while not satisfying $(A_1 \geq 30)$. Thus, some propositional constraints Σ forming a domain theory indicating how the Boolean conditions used in f are logically connected must be taken into account when computing explanations. Pairs (f, Σ) are referred to as *constrained decision-functions* in [Gorji and Rubin, 2022].

Because the feasible instances reduce to those satisfying Σ, leveraging Σ is mandatory to avoid the derivation of abductive explanations that are unnecessarily specific [Gorji and Rubin, 2022] or that could be simplified (for instance, the abductive explanation for $\boldsymbol{x}$ (associated with Alice) given by $\{(A_1 \geq 40), (A_1 \geq 30)\}$ can be simplified into $\{(A_1 \geq 40)\}$). It is also necessary to prevent from generating contrastive explanations that would correspond to instances that are impossible [Yu *et al.*, 2022], for example, the contrastive explanation for $\boldsymbol{x}'$ (associated with Bob) given by $\{\overline{(A_1 \geq 40)}\}$ that would correspond to the (impossible) contrastive instance given by $\{(A_1 \geq 40), (A_2 = 1), \overline{(A_1 \geq 30)}\}$.

In the following, our goal is to determine the computational impact of handling a domain theory Σ in the task of generating abductive explanations and contrastive explanations for instances given (f, Σ) when f is a decision tree. We consider the case when Σ is any theory, and also the more specific case when Σ is tractable. What we mean here by "tractable theory" Σ is the possibility of a polynomial-time clausal entailment tests: we suppose that a polynomial-time algorithm exists, that takes as input Σ and any clause δ, and returns true if and only if $\Sigma \models \delta$ holds.

Interestingly, knowledge compilation techniques can be exploited to "render tractable" propositional formulae Σ that are not tractable [Darwiche and Marquis, 2002]. Especially, there exist compilation algorithms associating with any CNF formula Σ an equivalent tractable theory (in the worst case, however, the resulting tractable theory is of size exponential in the size of Σ, so that no computational benefits can be guaranteed in every situation when knowledge compilation techniques are leveraged).

Among the tractable theories, we focus on two specific families, the Krom one (i.e., CNF formulae consisting of binary clauses) and the Horn one (i.e., CNF formulae where each clause contains at most one positive literal). It turns out that domain theories encoding numerical attributes or ordinal attributes are Krom theories. This is also the case of theories encoding categorical attributes (alias nominal attributes) under some open world assumption. Horn theories are also interesting because they can be used for encoding hierarchical features.

Our results are synthesized in Table 1. Each line of this table corresponds to a computation problem, that consists in deriving one (or all) explanations of a specific type for an input instance $\boldsymbol{x}$ given a constrained decision-function (f, Σ) where f is a decision tree. Each column corresponds to an assumption about the underlying theory Σ that is made. Each cell contains one of the following symbols: $\times$, $+$, or $\surd$. $\times$ means that the computation problem given by the line and the column is provably intractable, i.e., there is no polynomial-time algorithm to solve it. $+$ means that the computation problem given by the line and the column is probably intractable, i.e., there is no polynomial-time algorithm to solve the problem unless P = NP. Finally, $\surd$ indicates that the computation problem given by the line and the column is tractable, i.e., there exists a polynomial-time algorithm to solve the problem.

Table 1 clearly shows that taking advantage of Σ has a significant computational cost in the general case (just compare the two columns "Σ valid" and "any Σ").

Computation problem: deriving	Σ valid	any Σ	Σ tractable	Σ Horn	Σ Krom
One subset-minimal abductive explanation	$\surd$	$+$	$\surd$	$\surd$	$\surd$
All the subset-minimal abductive explanations	$\times$	$\times$	$\times$	$\times$	$\times$
One minimum-size abductive explanation	$+$	$+$	$+$	$+$	$+$
All the minimum-size abductive explanations	$\times$	$\times$	$\times$	$\times$	$\times$
One subset-minimal contrastive explanation	$\surd$	$+$	$\surd$	$\surd$	$\surd$
All the subset-minimal contrastive explanations	$\surd$	$\times$	$\times$	$\times$	$\surd$
One minimum-size contrastive explanation	$\surd$	$+$	$+$	$+$	$\surd$
All the minimum-size contrastive explanations	$\surd$	$\times$	$\times$	$\times$	$\surd$

Table 1: The complexity of deriving explanations given a constrained decision-function (f, Σ) when f is a decision tree. $\times$ means that the problem is provably intractable, $+$ means that the problem is intractable unless P = NP, and $\surd$ means that the problem is tractable.

2 Preliminaries

Classification Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a finite set of attributes, where each attribute is Boolean, categorical (aka nominal), or numerical. The *domain* D_i of A_i ($i \in [n]$) is $\{0, 1\}$ when A_i is Boolean, a finite set of values that are not ordered when A_i is categorical (for instance $D_i = \{blue, white, red\}$), and (typically) $D_i = \mathbb{N}$ or $\mathbb{R}$ when A_i is numerical. Note that the type of an attribute A_i is a semantical information that must be part of its description. Especially, it cannot be inferred from the values in the corresponding domain D_i (numbers can be used to denote values, like 0 for $blue$, 1 for $white$, and 2 for red, but it does not necessarily make sense in this case to consider that $0 < 1 < 2$). We note $\mathcal{A}_{boo}$ (resp. $\mathcal{A}_{num}$, $\mathcal{A}_{cat}$) the subset of $\mathcal{A}$ consisting of Boolean (resp. numerical, categorical) attributes.

An *instance* $\boldsymbol{x}$ over $\mathcal{A}$ is a vector from $D_1 \times \ldots \times D_n$. Every $\boldsymbol{x} = (v_1, \ldots, v_n)$ is also viewed logically as the conjunctively-interpreted set $t_{\boldsymbol{x}}$ of Boolean conditions (alias characteristics) $\{(A_i = v_i) : i \in [n]\}$. $\boldsymbol{X}$ is the set of all instances. A *binary classifier* f over $\mathcal{A}$ is a mapping from $\boldsymbol{X}$ to $\mathcal{L} = \{0, 1\}$. An instance $\boldsymbol{x} \in \boldsymbol{X}$ is *positive* when $f(\boldsymbol{x}) = 1$ and it is *negative* when $f(\boldsymbol{x}) = 0$.

A *decision tree* over $\mathcal{A}$ is a binary tree T, each of whose internal nodes is labeled with a Boolean condition on $A_i \in \mathcal{A}$, and each leaf is labeled by an element of $\mathcal{L}$. Without loss of generality, every variable is supposed to occur at most once on any root-to-leaf path. The value $T(\boldsymbol{x})$ of T on an input instance $\boldsymbol{x}$ is given by the label of the leaf reached from the root as follows: at each node go to the left (resp. right) child if the Boolean condition labelling the node is evaluated to 0 (resp. 1) for $\boldsymbol{x}$. The size of a decision tree is the number of nodes in it.

Boolean functions By $\mathcal{F}_n$ we denote the class of all Boolean functions from $\{0, 1\}^n$ to $\{0, 1\}$, and we use $X_n = \{x_1, \cdots, x_n\}$ to denote the set of input Boolean variables. A Boolean vector $\boldsymbol{x} \in \{0, 1\}^n$ is a *model* of f if $f(\boldsymbol{x}) = 1$. Otherwise, $\boldsymbol{x}$ is a *counter-model* of f. $[f]$ denotes the set of all models of f.

We refer to f as a propositional formula when it is described using the Boolean connectives $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation), together with the constants 1 (true) and 0 (false). f is *satisfiable* if it has a positive instance, and it is *unsatisfiable* otherwise. f is *valid* when it has no negative instance. If f and g are two propositional formulae over X_n, f *entails* g, noted $f \models g$, if and only if $[f] \subseteq [g]$ holds and f and g are *equivalent*, noted $f \equiv g$, if and only if $[f] = [g]$. A *literal* l_i is a variable $x_i \in X_n$ (a positive literal) or its negation $\neg x_i$ (a negative literal), also denoted $\overline{x}_i$. The complementary literal $\sim l_i$ of literal l_i is $\overline{x}_i$ if $l_i = x_i$ is a positive literal, and x_i if $l_i = \overline{x}_i$ is a negative literal. L_{X_n} is the set of all literals over X_n. A *term* t is a conjunction of literals, and a *clause* c is a disjunction of literals. In the following, we shall often treat instances as terms, and terms as sets of literals. A term t is an *implicant* of f if and only if $t \models f$ holds and t is a *prime implicant* of f if and only if t is an implicant of f and no proper subset of t is an implicant of f. A clause c is an *implicate* of f if and only if $f \models c$ holds, and c is a *prime implicate* of f if and only if c is an implicate of f and no proper subset of c is an implicate of f. A `DNF` *formula* is a disjunction of terms and a `CNF` *formula* is a conjunction of clauses. The set of variables occurring in a formula f is denoted $Var(f)$.

For an assignment $\boldsymbol{z} \in \{0, 1\}^n$, the corresponding canonical term is

$$t_{\boldsymbol{z}} = \bigwedge_{i=1}^{n} x_i^{z_i} \text{ where } x_i^0 = \overline{x}_i \text{ and } x_i^1 = x_i$$

A term t *covers* an assignment $\boldsymbol{x}$ if $t \subseteq t_{\boldsymbol{x}}$.

When every Boolean condition occurring in a decision tree T over a set $\mathcal{A}$ of attributes is viewed as a Boolean variable,

T can be viewed as a Boolean function over X_n. The class of decision trees over X_n is denoted $\mathtt{DT}_n$.

Finally, a *constrained decision-function* over a set of Boolean variables can be defined as follows [Gorji and Rubin, 2022]:

Definition 1. *Let $X_n = \{x_1, \ldots, x_n\}$ be a set of Boolean variables. A* constrained decision-function *over X_n is a pair (f, Σ) where $f \in F_n$ and Σ is a propositional formula over X_n. Σ indicates how the Boolean variables from X_n are logically connected.*

3 Explanations

Let us now define in formal terms the two types of explanations we are interested in: abductive explanations and contrastive explanations.

Definition 2. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance s.t. $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$).*

- *An* abductive explanation *for $\boldsymbol{x}$ given (f, Σ) is a set $t \subseteq t_{\boldsymbol{x}}$ such that $t \wedge \Sigma \models f$ (resp. $t \wedge \Sigma \models \overline{f}$).*
- *A* subset-minimal abductive explanation *for $\boldsymbol{x}$ given (f, Σ) is an abductive explanation t for $\boldsymbol{x}$ given (f, Σ) such that no proper subset of t is an abductive explanation for $\boldsymbol{x}$ given (f, Σ).*
- *A* minimum-size abductive explanation *for $\boldsymbol{x}$ given (f, Σ) is an abductive explanation t for $\boldsymbol{x}$ given (f, Σ) such that no abductive explanation t' for $\boldsymbol{x}$ given (f, Σ) such that $|t'| < |t|$ exists.*

Subset-minimal abductive explanations are also referred to as PI-explanations [Shih *et al.*, 2018], sufficient reasons [Darwiche and Hirth, 2020], or abductive explanations [Ignatiev *et al.*, 2019].

Definition 3. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance.*

- *A* contrastive explanation *for $\boldsymbol{x}$ given (f, Σ) is a set $c \subseteq t_{\boldsymbol{x}}$ such that the vector $\boldsymbol{x}_c \in \{0,1\}^n$ that coincides with $\boldsymbol{x}$ except on the characteristics of c (a so-called contrastive instance) is such that $\boldsymbol{x}_c \in [\Sigma]$ and $f(\boldsymbol{x}_c) \neq f(\boldsymbol{x})$.*
- *A* subset-minimal contrastive explanation *for $\boldsymbol{x}$ given (f, Σ) is a contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) such that no proper subset of c is a contrastive explanation for $\boldsymbol{x}$ given (f, Σ).*
- *A* minimum-size contrastive explanation *for $\boldsymbol{x}$ given (f, Σ) is a contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) such that no contrastive explanation c' for $\boldsymbol{x}$ given (f, Σ) such that $|c'| < |c|$ exists.*

Subset-minimal contrastive explanations are also referred to as necessary reasons [Darwiche and Ji, 2022] or contrastive explanations [Ignatiev *et al.*, 2020].

Clearly enough, every instance $\boldsymbol{x}$ has an abductive explanation given (f, Σ) that can be obtained without any computational effort, since $t_{\boldsymbol{x}}$ is such an explanation. Furthermore, provided that $\boldsymbol{x}$ is known, every contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) entirely defines a corresponding contrastive instance $\boldsymbol{x}_c$, and vice-versa, an instance $\boldsymbol{x}_c \in [\Sigma]$ such that $f(\boldsymbol{x}_c) \neq f(\boldsymbol{x})$ entirely defines a contrastive explanation c for $\boldsymbol{x}$ given (f, Σ). Finally, it is obvious that minimum-size abductive (resp. contrastive) explanations form a subset (in general, a proper subset) of the set of subset-minimal abductive (resp. contrastive) explanations.

In the following, we focus on the issue of deriving such explanations when f is a decision tree. We first recall known results for the case when no domain theory connecting the Boolean conditions that occur in f is available (or, equivalently, Σ is a valid formula). In such a case, it has been shown that:

- As to abductive explanations:
  - An instance $\boldsymbol{x}$ over X_n may have exponentially many abductive explanations given f, and even exponentially many subset-minimal abductive explanations, and exponentially many minimum-size abductive explanations in the number n of Boolean features [Audemard *et al.*, 2022b; 2022a].
  - Computing a subset-minimal abductive explanation for $\boldsymbol{x}$ given f can be done in time polynomial in the size of f and n [Izza *et al.*, 2020], but it is unlikely that we can enumerate subset-minimal abductive explanations for $\boldsymbol{x}$ given f in output polynomial time [de Colnet and Marquis, 2022].
  - Computing a minimum-size abductive explanation for $\boldsymbol{x}$ given f is NP-hard [Barceló *et al.*, 2020].
- As to contrastive explanations:
  - An instance $\boldsymbol{x}$ over X_n may have exponentially many contrastive explanations,[1] but only polynomially-many subset-minimal contrastive explanations in the number n of Boolean features [Audemard *et al.*, 2021b; Huang *et al.*, 2021].
  - Computing all the subset-minimal contrastive explanations for $\boldsymbol{x}$ given f can be done in time polynomial in the size of f and n [Audemard *et al.*, 2021b; Huang *et al.*, 2021].
  - As a direct consequence, computing all the minimum-size contrastive explanations for $\boldsymbol{x}$ given f can be done in time polynomial in the size of f and n.

4 The Impact of Domain Theories

Σ is any theory We first consider the case when Σ is any propositional formula. In such a case, the presence of Σ can make the derivation of some explanations computationally harder. Obviously, the case when no domain theory is available (i.e., Σ is valid) is a specific case of the general case (when Σ is unconstrained). As a consequence, all hardness results obtained for the case when no domain theory is available still hold in the general case:

- As to abductive explanations:
  - An instance $\boldsymbol{x}$ over X_n may have exponentially many abductive explanations given f and Σ, and

[1] Just because f may have exponentially many models and exponentially many counter-models.

even exponentially many minimum-size abductive explanations in the number n of Boolean features.
  - Computing a minimum-size abductive explanation for $\boldsymbol{x}$ given f and Σ is NP-hard.
- As to contrastive explanations:
  - An instance $\boldsymbol{x}$ over X_n may have exponentially many contrastive explanations.

Let us now look at the remaining issues.

Proposition 1. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a CNF formula, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. Computing a subset-minimal abductive explanation for $\boldsymbol{x}$ given (f, Σ) is NP-hard, even if f reduces to a stump.*

Proof. Suppose that $f(\boldsymbol{x}) = 1$. By construction, the empty term $\top$ is the unique subset-minimal abductive explanation for $\boldsymbol{x}$ given (f, Σ) if and only if $\Sigma \models f$ holds. Similarly, if $f(\boldsymbol{x}) = 0$, then $\top$ is the unique subset-minimal abductive explanation for $\boldsymbol{x}$ given (f, Σ) if and only if $\Sigma \models \overline{f}$ holds. Thus, computing in (deterministic) polynomial time a subset-minimal abductive explanation for $\boldsymbol{x}$ when $\boldsymbol{x}$ is a positive instance is sufficient to decide in (deterministic) polynomial time whether $\Sigma \models f$. We show that this decision problem is coNP-hard, by reduction from CNF-UNSAT. Let $\alpha = \bigwedge_{i=1}^{m} \delta_i$ be a CNF formula over $\{x_1, \ldots, x_n\}$. We associate with α in polynomial time the CNF formula $\Sigma = \bigwedge_{i=1}^{m} (\delta_i \vee x_{n+1})$ where x_{n+1} is a fresh variable, f is a decision tree equivalent to x_{n+1} (it is a stump: f has a single internal node labelled by x_{n+1}, its left child is a 0-leaf and its right child is a 1-leaf), and $\boldsymbol{x}$ is the instance over $\{x_1, \ldots, x_n, x_{n+1}\}$ where every variable x_i ($i \in [n+1]$) is set to 1. We can easily check that $\boldsymbol{x}$ is a model of $\Sigma \wedge f$. Now, $\Sigma \models f$ holds if and only if $\overline{\alpha} \vee x_{n+1}$ is valid if and only if α is unsatisfiable. □

As a consequence, subset-minimal abductive explanations cannot be enumerated in output polynomial time unless P = NP.

Similarly, for subset-minimal contrastive explanations, we have that:

Proposition 2. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a CNF formula, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. Computing a subset-minimal contrastive explanation for $\boldsymbol{x}$ given (f, Σ) (or just a contrastive explanation for $\boldsymbol{x}$ given (f, Σ)) is NP-hard, and this holds even if f reduces to a stump.*

Proof. Towards a contradiction, suppose that computing a subset-minimal or even a contrastive explanation for $\boldsymbol{x}$ given (f, Σ) is feasible in (deterministic) polynomial time. Then one would be able to decide in (deterministic) polynomial time whether a contrastive explanation for $\boldsymbol{x}$ given (f, Σ) exists. However, this problem is NP-hard, as shown by the following reduction from CNF-SAT. Let $\alpha = \bigwedge_{i=1}^{m} \delta_i$ be a CNF formula over $\{x_1, \ldots, x_n\}$. We associate with α in polynomial time the CNF formula $\Sigma = \bigwedge_{i=1}^{m} (\delta_i \vee x_{n+1})$ where x_{n+1} is a fresh variable, f is a decision tree equivalent to x_{n+1} (it is a stump: f has a single internal node labelled by x_{n+1}, its left child is a 0-leaf and its right child is a 1-leaf), and $\boldsymbol{x}$ is the instance over $\{x_1, \ldots, x_n, x_{n+1}\}$ where every variable x_i ($i \in [n+1]$) is set to 1. We can easily check that $\boldsymbol{x}$ is a model of $\Sigma \wedge f$. So $\boldsymbol{x}$ has a contrastive explanation given (f, Σ) if and only if $\Sigma \wedge \overline{f}$ is satisfiable. But $\Sigma \wedge \overline{f}$ is equivalent to $\alpha \wedge \overline{x_{n+1}}$, which is satisfiable if and only if α is satisfiable. □

Unless P = NP, the previous result prevents from the polynomial-time generation of all subset-minimal contrastive explanations for $\boldsymbol{x}$ given (f, Σ), which is feasible when Σ is valid. Actually, the result can be proved *unconditionally* due to the number of subset-minimal contrastive explanations for $\boldsymbol{x}$ given (f, Σ). Indeed, it can be the case that the *minimum-size* contrastive explanations for $\boldsymbol{x}$ given (f, Σ) are exponentially numerous in the number of features used, and this holds not only in the general case when Σ is any theory, but also in the specific case when Σ is a tractable theory. Indeed:

Proposition 3. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a CNF formula, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. The number of minimum-size contrastive explanations for $\boldsymbol{x}$ given (f, Σ) can be exponential in the number of Boolean variables used in (f, Σ), and this is the case even when Σ is a Horn CNF formula or a CNF formula representing a set of domain constraints for categorical attributes where each domain contains at least 3 elements.*

Proof. Let us start with the case when Σ is a Horn CNF formula. Let us take $\Sigma = \bigwedge_{i=1}^{n} (\overline{x_i} \vee \overline{y_i} \vee z_i)$. Let f be a decision tree equivalent to the clause $\bigvee_{i=1}^{n} z_i$ (such a decision tree can be generated in time linear in n), and let $\boldsymbol{x}$ be the assignment over $\{x_i, y_i, z_i : i \in [n]\}$ where every variable is set to 1. The instance $\boldsymbol{x}$ is a model of $\Sigma \wedge f$. In order to get a contrastive instance $\boldsymbol{x}'$ satisfying $\Sigma \wedge \overline{f}$, the truth value of every variable z_i ($i \in [n]$) must be set to 0 since $\overline{f} \equiv \bigwedge_{i=1}^{n} \overline{z_i}$. But the assignment that coincides with $\boldsymbol{x}$ except on the variables z_i ($i \in [n]$) is not a model of Σ: it violates every clause in it. In order to make it a model of clause $\overline{x_i} \vee \overline{y_i} \vee z_i$ ($i \in [n]$) while keeping z_i set to 0, at least one of x_i or y_i must be set to 1. Thus there are 2^n to minimally change $\boldsymbol{x}$ to get a model $\boldsymbol{x}'$ of $\Sigma \wedge \overline{f}$: every set containing all the variables z_i ($i \in [n]$) and for each $i \in [n]$, precisely one of x_i or y_i is a minimum-size contrastive explanation for $\boldsymbol{x}$ given (f, Σ). The number of such sets is equal to 2^n, thus exponential in the number $3n$ of Boolean variables used in (f, Σ).

Let us now turn to the case when Σ is a CNF formula representing a set of domain constraints for categorical attributes where each domain contains at least 3 elements. Let us take $\Sigma = \bigwedge_{i=1}^{n} (x_i \vee y_i \vee z_i) \wedge (\overline{x_i} \vee \overline{y_i}) \wedge (\overline{x_i} \vee \overline{z_i}) \wedge (\overline{y_i} \vee \overline{z_i})$. Σ is a propositional encoding of domain constraints for n categorical attributes, where each domain contains 3 elements. Such a Σ is tractable since it is equivalent to the conjunction of its prime implicates. Let f be a decision tree equivalent to the clause $\bigvee_{i=1}^{n} z_i$ (such a decision tree can be generated in time linear in n), and let $\boldsymbol{x}$ be the assignment over $\{x_i, y_i, z_i : i \in [n]\}$ where every variable z_i ($i \in [n]$) is set to 1 and every variable x_i, y_i ($i \in [n]$) is set to 0 . Then the rest of the proof is similar to the Horn case above. The instance $\boldsymbol{x}$ is a model of $\Sigma \wedge f$. In order to get a contrastive instance $\boldsymbol{x}'$ satisfying $\Sigma \wedge \overline{f}$, the truth value of every variable z_i ($i \in [n]$)

must be set to 0 since $\overline{f} \equiv \bigwedge_{i=1}^{n} \overline{z_i}$. But the assignment that coincides with $\boldsymbol{x}$ except on the variables z_i ($i \in [n]$) is not a model of Σ: it violates every clause $x_i \vee y_i \vee z_i$ ($i \in [n]$) in it. In order to make it a model of Σ while keeping z_i set to 0, precisely one of x_i or y_i must be set to 1. Thus there are 2^n ways to minimally change $\boldsymbol{x}$ to get a model $\boldsymbol{x}'$ of $\Sigma \wedge \overline{f}$: every set containing all the variables z_i ($i \in [n]$) and for each $i \in [n]$, precisely one of x_i or y_i is a minimum-size contrastive explanation for $\boldsymbol{x}$ given (f, Σ). The number of such sets is equal to 2^n, thus exponential in the number $3n$ of Boolean variables used in (f, Σ). □

Σ is a tractable theory Let us now consider the case when Σ is a tractable theory. It turns out that supposing that Σ is tractable (implicitly, for the clausal entailment task) changes the picture when it comes to derive a subset-minimal abductive explanation:

Proposition 4. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a tractable theory, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. Computing a subset-minimal abductive explanation for $\boldsymbol{x}$ given (f, Σ) can be done in time polynomial in the size of the input.*

Proof. In order to generate efficiently a subset-minimal abductive explanation t for $\boldsymbol{x}$ given (f, Σ), one takes advantage of a greedy algorithm. If $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$), then we turn f (resp. $\overline{f}$) into an equivalent CNF formula $\bigwedge_{i=1}^{m} \delta_i$ (it is well-known that this can be achieved in time linear in the size of f). Then we initialize t to $t_{\boldsymbol{x}}$, and consider every literal ℓ of t in sequence (the order chosen does not matter to prove the result). At each step, an implicant test is performed: one tests whether $(t \setminus \{\ell\}) \wedge \Sigma \models \bigwedge_{i=1}^{m} \delta_i$. If the test succeeds, then ℓ is removed from t, else it is kept, and the algorithm resumes considering the next literal of t in the sequence. When every literal has been considered, the resulting term t is by construction a subset-minimal abductive explanation t for $\boldsymbol{x}$ given (f, Σ). Each test $(t \setminus \{\ell\}) \wedge \Sigma \models \bigwedge_{i=1}^{m} \delta_i$ can be done in polynomial time. Indeed, the condition holds precisely when for every clause δ_i ($i \in [m]$), $\Sigma \models \neg(t \setminus \{\ell\}) \vee \delta_i$. Since $\neg(t \setminus \{\ell\}) \vee \delta_i$ is a clause and Σ is a tractable theory, the condition can be evaluated in polynomial time. Since the number of implicant tests to be achieved is equal to the number of features, the greedy algorithm runs in time polynomial in the size of the input. □

Focusing now on the generation of subset-minimal contrastive explanations, it is valuable to consider a further restriction on the tractable theory at hand, namely that Σ is a Krom CNF formula (i.e., Σ is given as a conjunction of binary clauses). Such theories are known as tractable for a while [Even *et al.*, 1976; Aspvall *et al.*, 1979]. Indeed, when Σ is a Krom CNF formula, the computation of all the subset-minimal contrastive explanations for $\boldsymbol{x}$ given (f, Σ) can be done in time polynomial in the size of the input. As a direct consequence, the computation of all the minimum-size contrastive explanations for $\boldsymbol{x}$ given (f, Σ) can also be achieved in time polynomial in the size of the input. So in the case when Σ is a Krom CNF formula, the results obtained for the case when Σ is valid still hold.

Proposition 5. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a Krom CNF formula, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. Computing all the subset-minimal contrastive explanations for $\boldsymbol{x}$ given (f, Σ) can be done in time polynomial in the size of the input.*

Proof. First of all, if $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$), then we turn $\overline{f}$ (resp. f) into an equivalent DNF formula $\bigvee_{i=1}^{m} \gamma_i$ (it is well-known that this can be achieved in time linear in the size of f). Then the goal is to determine models $\boldsymbol{x}_c$ of $\Sigma \wedge \bigvee_{i=1}^{m} \gamma_i$ (i.e., contrastive instances) such that the set-theoretic difference of $t_{\boldsymbol{x}}$ minus $t_{\boldsymbol{x}_c}$ is minimal w.r.t. set-inclusion. Now, $\Sigma \wedge \bigvee_{i=1}^{m} \gamma_i$ is equivalent to $\bigvee_{i=1}^{m} (\Sigma \wedge \gamma_i)$. This shows that to get the models $\boldsymbol{x}_c$ we look for, we can look at the models of each $\Sigma \wedge \gamma_i$ ($i \in [m]$). Every γ_i such that $\Sigma \wedge \gamma_i$ is unsatisfiable can be detected in polynomial time since Σ is tractable and $\Sigma \wedge \gamma_i$ is unsatisfiable if and only if $\Sigma \models \neg\gamma_i$ where $\neg\gamma_i$ is a clause. When $\Sigma \wedge \gamma_i$ is unsatisfiable, γ_i can be set aside since no model $\boldsymbol{x}_c$ among the ones we look for can be a model of $\Sigma \wedge \gamma_i$ because $\Sigma \wedge \gamma_i$ has no model. If this happens for every $i \in [m]$, then $\Sigma \wedge \bigvee_{i=1}^{m} \gamma_i$ has no model, so that no contrastive explanation exists.

The next step is to prove that for every γ_i that is remaining, there exists a *unique* model $\boldsymbol{x}_{c_i}$ of $\Sigma \wedge \gamma_i$ such that $t_{\boldsymbol{x}}$ minus $t_{\boldsymbol{x}_{c_i}}$ is minimal w.r.t. set-inclusion. Clearly enough, every candidate $\boldsymbol{x}_{c_i}$ must satisfy γ_i so the truth value of every literal ℓ of $t_{\boldsymbol{x}}$ such that $\overline{\ell} \in \gamma_i$ must be switched in $\boldsymbol{x}_{c_i}$, while the truth value of every literal ℓ of $t_{\boldsymbol{x}}$ such that $\ell \in \gamma_i$ is kept in $\boldsymbol{x}_{c_i}$ as it is in $t_{\boldsymbol{x}}$. More generally, every literal ℓ such that $\Sigma \wedge \gamma_i \models \ell$ must be satisfied by $\boldsymbol{x}_{c_i}$ since $\boldsymbol{x}_{c_i}$ must be a model of $\Sigma \wedge \gamma_i$. The set L_i of literals ℓ entailed by $\Sigma \wedge \gamma_i$ can be computed in polynomial time because Σ is tractable ($\Sigma \wedge \gamma_i \models \ell$ holds if and only if the clause $\neg\gamma_i \vee \ell$ is entailed by Σ). So let $\boldsymbol{x}_{c_i}$ be defined as the assignment that satisfies every literal from L_i and that coincides with $\boldsymbol{x}$ on every variable x so that neither x nor $\overline{x}$ belongs to L_i. By construction, all the models of $\Sigma \wedge \gamma_i$ give the same truth values to the literals in L_i. Hence $t_{\boldsymbol{x}}$ minus $t_{\boldsymbol{x}_{c_i}}$ is minimal w.r.t. set-inclusion. Therefore, it remains to show that $\boldsymbol{x}_{c_i}$ is a model of $\Sigma \wedge \gamma_i$. Since it is a model of γ_i, one just has to show that $\boldsymbol{x}_{c_i}$ is a model of Σ. Consider any clause δ of Σ. There are two cases. (1) If δ contains at least one literal $\ell \in L_i$, then δ is satisfied by $\boldsymbol{x}_{c_i}$ since $\boldsymbol{x}_{c_i}$ sets ℓ to true. (2) Else, no literal of $\delta = \ell_1 \vee \ell_2$ belongs to L_i. In this case, neither $\overline{\ell_1}$ nor $\overline{\ell_2}$ belongs to L_i as well, otherwise by unit propagation we would have been in case (1) or $\Sigma \wedge \gamma_i$ would have been unsatisfiable. So the variables of ℓ_1 and ℓ_2 are set in $\boldsymbol{x}_{c_i}$ to the same truth values as the ones they have in $\boldsymbol{x}$. But since $\boldsymbol{x}$ satisfies Σ, $\boldsymbol{x}$ satisfies δ, hence $\boldsymbol{x}_{c_i}$ satisfies δ as well.

□

Notably, the theories Σ obtained by encoding numerical and/or ordinal attributes are Krom CNF formulae. This is also the case of theories encoding categorical attributes, provided that an open world assumption is made. What we mean here is that if $V = \{v_1, \ldots, v_p\}$ is the set of values of a categorical attribute A such that nodes of the type $(A = v_i)$ ($i \in [p]$) are encountered in the decision tree f, then Σ is equivalent to the Krom CNF formula $\bigwedge_{v_j, v_k \in V | v_i \neq v_k} (\overline{(A = v_j)} \vee \overline{(A = v_k)})$,

i.e., the values in V are mutually exclusive. Thus, the constraint $\bigvee_{v_j \in V} v_j$ is not implied by Σ: other values than those listed in V are considered as possible for attribute A. So when considering decision trees f based on numerical and/or ordinal features and/or categorical features under an open world assumption, the generation of all the subset-minimal contrastive explanations for an instance $\boldsymbol{x}$ given (f, Σ) can be achieved in polynomial time. And, as a consequence, the generation of all the minimum-size contrastive explanations for an instance $\boldsymbol{x}$ given (f, Σ) can be achieved in polynomial time as well.

We now focus on domain theories Σ that are tractable but do not reduce to Krom CNF formulae. Among them are (for instance) Horn CNF formulae. Such Horn theories are tractable and they are interesting because they can be used for encoding hierarchical features, for instance the fact that every plane geometry object that satisfies the property "rectangle" and the property "diamond" must have the property "square" as well. Because of Proposition 3, we already know that the generation of all subset-minimal contrastive explanations for an instance $\boldsymbol{x}$ given (f, Σ) cannot be achieved in polynomial time when Σ is a Horn CNF formula (so the result extends to tractable theories in the general case). Thus, we need to focus on computationally easier problems, namely the generation of one subset-minimal contrastive explanation and the generation of one minimum-size contrastive explanation.

We have obtained the following results:

Proposition 6. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a tractable theory, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. Computing one subset-minimal contrastive explanation for $\boldsymbol{x}$ given (f, Σ) can be done in time polynomial in the size of the input.*

Proof. The proof goes through a number of intermediate results.

The first step of the proof is similar to the first step in the proof of Proposition 5. If $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$), then we turn $\overline{f}$ (resp. f) in linear time into an equivalent DNF formula $\bigvee_{i=1}^{m} \gamma_i$. Then one looks for a model $\boldsymbol{x}_c$ of $\Sigma \wedge \bigvee_{i=1}^{m} \gamma_i$ such that the $t_{\boldsymbol{x}}$ minus $t_{\boldsymbol{x}_c}$ is minimal w.r.t. set-inclusion. Since $\Sigma \wedge \bigvee_{i=1}^{m} \gamma_i$ is equivalent to $\bigvee_{i=1}^{m} (\Sigma \wedge \gamma_i)$, for each $i \in [m]$, we look first to a model $\boldsymbol{x}_c^i$ of $\Sigma \wedge \gamma_i$ such that $t_{\boldsymbol{x}}$ minus $t_{\boldsymbol{x}_c^i}$ is minimal w.r.t. set-inclusion (among the models of $\Sigma \wedge \gamma_i$).

Given two instances $\boldsymbol{x}, \boldsymbol{x}' \in \{0,1\}^n$, we define $\boldsymbol{x} \oplus \boldsymbol{x}'$ as the element of $\{0,1\}^n$ such that for each $j \in [n]$, $(\boldsymbol{x} \oplus \boldsymbol{x}')_j = 1$ if $\boldsymbol{x}_i \neq \boldsymbol{x}'_i$ and $(\boldsymbol{x} \oplus \boldsymbol{x}')_j = 0$ if $\boldsymbol{x}_i = \boldsymbol{x}'_i$. Observe that $\boldsymbol{x} \oplus \boldsymbol{x}'$ is totally defined from $\boldsymbol{x}$ and $\boldsymbol{x}'$, and that conversely, $\boldsymbol{x}'$ is totally defined from $\boldsymbol{x}$ and $\boldsymbol{x} \oplus \boldsymbol{x}'$ (we have $\boldsymbol{x}' = \boldsymbol{x} \oplus (\boldsymbol{x} \oplus \boldsymbol{x}')$). Stated otherwise, given $\boldsymbol{x}$, $\boldsymbol{x}'$ and $\boldsymbol{x} \oplus \boldsymbol{x}'$ are in one-to-one correspondance.

We consider two orderings over $\{0,1\}^n$: the product ordering $\leq_{prod}$ induced by $0 < 1$ (it is a partial ordering) and the lexicographic ordering $\leq_{lex}$ over $\{0,1\}^n$ induced by $0 < 1$ (it is a total ordering). It is well-known that $\leq_{lex}$ is a linear extension of $\leq_{prod}$, meaning that for any $\boldsymbol{x}', \boldsymbol{x}'' \in \{0,1\}^n$, if $\boldsymbol{x}' \leq_{prod} \boldsymbol{x}''$ holds, then $\boldsymbol{x}' \leq_{lex} \boldsymbol{x}''$ holds.

For any $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{x}'' \in \{0,1\}^n$ we note $\boldsymbol{x}' \sqsubseteq_{\boldsymbol{x}} \boldsymbol{x}''$ precisely when $\boldsymbol{x} \oplus \boldsymbol{x}' \leq_{prod} \boldsymbol{x} \oplus \boldsymbol{x}''$, and $\boldsymbol{x}' \preceq_{\boldsymbol{x}} \boldsymbol{x}''$ precisely when $\boldsymbol{x} \oplus \boldsymbol{x}' \leq_{lex} \boldsymbol{x} \oplus \boldsymbol{x}''$. By construction, we have $\boldsymbol{x}' \sqsubseteq_{\boldsymbol{x}} \boldsymbol{x}''$ holds if and only $t_{\boldsymbol{x}} \setminus t_{\boldsymbol{x}'} \subseteq t_{\boldsymbol{x}} \setminus t_{\boldsymbol{x}''}$ holds.

Now, since $\leq_{lex}$ is a total ordering, for each $i \in [m]$, the set $\{\boldsymbol{x} \oplus \boldsymbol{x}' : \boldsymbol{x}' \in [\Sigma \wedge \gamma_i]\}$ has a least element w.r.t. $\leq_{lex}$. Let $\boldsymbol{x} \oplus \boldsymbol{x}_c^i$ denote this element. By construction, $\boldsymbol{x}_c^i$ is the least element of $[\Sigma \wedge \gamma_i]$ w.r.t. $\preceq_{\boldsymbol{x}}$.

In order to compute $\boldsymbol{x}_c^i$ we can take advantage of the following algorithm:

Algorithm 1 Computing $\boldsymbol{x}_c^i$

Require: a tractable theory Σ, a term γ_i such that $\Sigma \not\models \neg\gamma_i$, and an instance $\boldsymbol{x} = (\ell_1, \ldots, \ell_n) \in [\Sigma \wedge \neg\gamma_i]$
Ensure: $t_{\boldsymbol{x}_c^i}$, where $\boldsymbol{x}_c^i$ is the least element of $[\Sigma \wedge \gamma_i]$ w.r.t. $\preceq_{\boldsymbol{x}}$

```
t ← γ_i
for i = 1 to n do
    if Σ ⊭ ¬t ∨ ¬ℓ_i then
        t ← t ∧ ℓ_i
    else
        t ← t ∧ ∼ℓ_i
    end if
end for
return (t)
```

This algorithm runs in polynomial time whenever Σ is tractable for clausal entailment. Basically, the term t that is computed by this algorithm implies γ_i and every literal of $t_{\boldsymbol{x}}$ is conjoined with t in sequence, provided that it does not conflict with Σ once augmented by the literals of $t_{\boldsymbol{x}}$ that have been previously conjoined (note that we could simplify the loop and avoid testing whether $\Sigma \not\models \neg t \vee \neg\ell_i$ whenever $\ell_i \in \gamma_i$, but we keep the pseudo-code as it is for the sake of readability).

It turns out that the contrastive explanation c for $\boldsymbol{x}$ given (γ_i, Σ) associated with $\boldsymbol{x}_c^i$ and given by $c = t_{\boldsymbol{x}} \setminus t_{\boldsymbol{x}_c^i}$ is subset-minimal. Indeed, if it was not the case, there would exist a subset-minimal contrastive explanation c' for $\boldsymbol{x}$ given (γ_i, Σ) such that $c' \subset c$. Let $\boldsymbol{x}_{c'}^i$ be the corresponding contrastive instance. We would have $\boldsymbol{x}_{c'}^i \sqsubset_{\boldsymbol{x}} \boldsymbol{x}_c^i$ and as a consequence (since $\preceq_{\boldsymbol{x}}$ extends $\sqsubset_{\boldsymbol{x}}$), we would also have $\boldsymbol{x}_{c'}^i \prec_{\boldsymbol{x}} \boldsymbol{x}_c^i$, contradicting the fact that $\boldsymbol{x}_c^i$ is the least element of $[\Sigma \wedge \gamma_i]$ w.r.t. $\preceq_{\boldsymbol{x}}$.

Finally, we compute in polynomial time the least element noted $\boldsymbol{x}_c$ of $\{\boldsymbol{x}_c^i : i \in [m]\}$ w.r.t. $\preceq_{\boldsymbol{x}}$. This element is the least element of $[\Sigma \wedge \bigvee_{i=1}^{m} \gamma_i]$ w.r.t. $\preceq_{\boldsymbol{x}}$. Consequently, the corresponding $c = t_{\boldsymbol{x}} \setminus t_{\boldsymbol{x}_c}$ is a subset-minimal contrastive explanation for $\boldsymbol{x}$ given (f, Σ). □

Proposition 7. *Let (f, Σ) be a constrained decision-function, where f is a decision tree, Σ is a tractable theory, and let $\boldsymbol{x} \in [\Sigma]$ be an instance. Computing one minimum-size contrastive explanation for $\boldsymbol{x}$ given (f, Σ) is* NP*-hard and this holds even if Σ is a pure Horn* CNF *formula and f is a stump.*

Proof. Suppose that a (deterministic) polynomial-time algorithm for computing one minimum-size contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) exists. If $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$),

then $|c|$ is the (minimal) Hamming distance between $\boldsymbol{x}$ and $[\Sigma \wedge \overline{f}]$ (resp. $[\Sigma \wedge f]$). Accordingly, if such an algorithm existed, one would be able to decide in (deterministic) polynomial-time whether the Hamming distance between a model $\boldsymbol{x}$ of $\Sigma \wedge f$ and $[\Sigma \wedge \overline{f}]$ is lower than or equal to any given non-negative integer, say d'.

We now show that the latter problem is NP-hard by reducing the well-known minimal hitting set problem (MHS) to it. An instance of MHS is given by a pair (C, d) where C is a finite set of subsets of a finite set $S = \{x_1, \ldots, x_n\}$ and d is a non-negative integer; the instance is positive if and only if there exists a subset H of S s.t. $|H| \leq d$ and H is a hitting set of C, i.e., for every $c \in C$, $H \cap c \neq \varnothing$. It is known that MHS is NP-complete even if the case when each $c \in C$ contains at most two elements [Karp, 1972] (which is an assumption we make here). With such an instance (C, d), we associate in polynomial time the pure Horn formula $\Sigma = \bigwedge_{c \in C}(\bigvee_{x_i \in c} \overline{x_i}) \vee x_0$ over $\{x_0, x_1, \ldots, x_n\}$, the stump f equivalent to x_0, the instance $\boldsymbol{x}$ such that $\forall j \in \{0, \ldots, n\}$, $\boldsymbol{x}_j = 1$, and $d' = d + 1$. We can easily check that $\boldsymbol{x}$ is a model of $\Sigma \wedge f$. Furthermore, $\Sigma \wedge \overline{f}$ is equivalent to $\bigwedge_{c \in C}(\bigvee_{x_i \in c} \overline{x_i}) \wedge \overline{x_0}$. Since x_0 is set to 1 in $\boldsymbol{x}$ and x_0 does not occur in S, the Hamming distance between $\boldsymbol{x}$ and $\Sigma \wedge \overline{f}$ is equal to 1 plus the Hamming distance between $\boldsymbol{x}$ and $\bigwedge_{c \in C}(\bigvee_{x_i \in c} \overline{x_i})$. Since every variable is set to 1 in $\boldsymbol{x}$, updating $\boldsymbol{x}$ so as to satisfy a clause $\bigvee_{x_i \in c} \overline{x_i}$ where $c \in C$ requires to set to 0 in $\boldsymbol{x}$ at least one of the (at most two) variables x_i. Since $\bigwedge_{c \in C}(\bigvee_{x_i \in c} \overline{x_i})$ must be satisfied, a minimal change in terms of the number of variables to be flipped in $\boldsymbol{x}$ to get a model of $\bigwedge_{c \in C}(\bigvee_{x_i \in c} \overline{x_i})$ (i.e., the Hamming distance between $\boldsymbol{x}$ and $\bigwedge_{c \in C}(\bigvee_{x_i \in c} \overline{x_i})$) is given by the size of a minimal hitting set of C. Thus, C has a hitting set of size $\leq d$ if and only if a contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) of size $\leq d + 1$ exists. This concludes the proof. □

5 Conclusion

In this paper, we have shown that leveraging a domain theory indicating how the Boolean conditions occurring in a decision tree are logically connected may have a strong impact on the complexity of generating provably correct explanations. None of the explanation problems that are tractable when the Boolean conditions used in the tree are independent (i.e., Σ is valid) remain tractable when no assumptions are made on the domain theory. Ensuring that Σ is tractable is enough to preserve the results about the computation of abductive explanations that hold when Σ is valid, but in general, it changes significantly the picture concerning the computation of contrastive explanations (the sole contrastive explanation problem that is tractable when Σ is valid and that remains tractable when Σ is a tractable domain theory is the computation of one subset-minimal contrastive explanation). Contrastingly, when Σ is a Krom CNF formula, all the explanation problems that are tractable when Σ is valid remain tractable. The practical significance of this result comes notably from the fact that the domain theories Σ obtained by encoding numerical and/or ordinal attributes are Krom CNF formulae.

Acknowledgements

This work has benefited from the support of the AI Chair EXPEKCTATION (ANR-19-CHIA-0005-01) of the French National Research Agency (ANR). It was also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[Aspvall *et al.*, 1979] B. Aspvall, M. Plass, and R. Tarjan. A linear-time algorithm for testing the truth of certain quantified Boolean formulas. *Information Processing Letters*, 8:121–123, 1979. Erratum: Information Processing Letters 14(4): 195 (1982).

[Audemard *et al.*, 2021a] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021*, pages 74–86, 2021.

[Audemard *et al.*, 2021b] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of decision trees. *CoRR*, abs/2108.05266, 2021.

[Audemard *et al.*, 2022a] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI'22*, pages 643–650, 2022.

[Audemard *et al.*, 2022b] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of boolean decision trees. *Data Knowl. Eng.*, 142:102088, 2022.

[Audemard *et al.*, 2023] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. On contrastive explanations for tree-based classifiers. In *Proc. of ECAI'23*, 2023. To appear.

[Barceló *et al.*, 2020] P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux. Model interpretability through the lens of computational complexity. In *Proc. of NeurIPS'20*, 2020.

[Breiman *et al.*, 1984] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[Breiman, 2001] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[Darwiche and Hirth, 2020] A. Darwiche and A. Hirth. On the reasons behind decisions. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI'20)*, pages 712–720, 2020.

[Darwiche and Ji, 2022] Adnan Darwiche and Chunxi Ji. On the computation of necessary and sufficient explanations. In *Proc. of AAAI'22*, pages 5582–5591, 2022.

[Darwiche and Marquis, 2002] A. Darwiche and P. Marquis. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17:229–264, 2002.

[de Colnet and Marquis, 2022] Alexis de Colnet and Pierre Marquis. On the complexity of enumerating prime implicants from decision-dnnf circuits. In *Proc. of IJCAI'22*, pages 2583–2590, 2022.

[Even *et al.*, 1976] S. Even, A. Itai, and A. Shamir. On the complexity of timetable and integral multi-commodity flow problems. *SIAM J. on Comp*, 5(4):691–703, 1976.

[Freund and Schapire, 1997] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[Friedman, 2001] J. H. Friedman. Greedy function approximation: A gradient boosted machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[Gorji and Rubin, 2022] Niku Gorji and Sasha Rubin. Sufficient reasons for classifier decisions in the presence of domain constraints. In *Proc. of AAAI'22*, pages 5660–5667, 2022.

[Huang *et al.*, 2021] X. Huang, Y. Izza, A. Ignatiev, and J. Marques-Silva. On efficiently explaining graph-based classifiers. In *Proc. of KR'21*, pages 356–367, 2021.

[Ignatiev *et al.*, 2019] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of AAAI'19*, pages 1511–1519, 2019.

[Ignatiev *et al.*, 2020] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva. From contrastive to abductive explanations and back again. In *AIxIA 2020 - Advances in Artificial Intelligence - XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25-27, 2020, Revised Selected Papers*, volume 12414 of *Lecture Notes in Computer Science*, pages 335–355, 2020.

[Izza *et al.*, 2020] Y. Izza, A. Ignatiev, and J. Marques-Silva. On explaining decision trees. *CoRR*, abs/2010.11034, 2020.

[Karp, 1972] R.M. Karp. *Reducibility among combinatorial problems*, chapter Complexity of Computer Computations, pages 85–103. Plenum Press, New York, 1972.

[Miller, 2019] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Quinlan, 1986] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[Ras *et al.*, 2022] Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *J. Artif. Intell. Res.*, 73:329–396, 2022.

[Schapire and Freund, 2014] R.E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2014.

[Shih *et al.*, 2018] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI'18)*, pages 5103–5111, 2018.

[Yu *et al.*, 2022] Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and João Marques-Silva. Eliminating the impossible, whatever remains must be true. *CoRR*, abs/2206.09551, 2022.