ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230305

Abductive Explanations of Classifiers Under Constraints: Complexity and Properties

Martin Cooper[a;*] **and Leila Amgoud**[b]

[a]University of Toulouse 3, France
[b]CNRS, IRIT France
ORCiD ID: Martin Cooper https://orcid.org/0000-0003-4853-053X,
Leila Amgoud https://orcid.org/0000-0002-1838-4271

Abstract. Abductive explanations (AXp's) are widely used for understanding decisions of classifiers. Existing definitions are suitable when features are *independent*. However, we show that ignoring constraints when they exist between features may lead to an explosion in the number of redundant or superfluous AXp's. We propose three new types of explanations that take into account constraints and that can be generated from the whole feature space or from a sample (such as a dataset). They are based on a key notion of coverage of an explanation, the set of instances it explains. We show that coverage is powerful enough to discard redundant and superfluous AXp's. For each type, we analyse the complexity of finding an explanation and investigate its formal properties. The final result is a catalogue of different forms of AXp's with different complexities and different formal guarantees.

1 Introduction

Given a decision of a classifier, a user may want, and may even have a legal right to, an explanation of this decision. Concrete examples include an explanation as to why a loan/job/visa application was refused or why a medical diagnosis was made. See [21, 22] for more on explainability and interpretability.

The majority of existing explanation functions explain a decision in terms of relevance of the input features. One of the most studied types of feature-based explanations is the so-called *abductive explanation* (AXp), or *prime implicant explanation* [10, 19, 25]. It provides a (minimal) sufficient reason for the decision.

In the literature, AXp's have two sources: they are generated either from a subset of instances as done by the two prominent explanation functions Anchors [24] and LIME [23] and those introduced in [1, 3], or from the whole feature space (eg., [4, 5, 6, 8, 10, 11, 14, 15, 25]). Whatever the source, features are *implicitly* assumed to be *independent*. However, constraints on values that features may take are ubiquitous in almost all real-world applications including analysis of election results, justifying medical treatments, etc.

Constraints have been extensively studied in databases where several types have been distinguished [26]. In the context of classifiers and their abductive explanations, we focus on two categories: *integrity constraints* (IC) and *dependency constraints* (DC). The former are of two types: i) they may express impossible assignments of values to features like "men cannot be pregnant", here ICs impact locally individual instances, ii) global constraints preventing the coexistence of two or more instances such as "no two distinct students may have the same ID card value". When such constraints exist, the feature space necessarily contains *impossible instances*. Dependency constraints are a specific sub-type of the first type of IC. They express the following: if some attributes take specific values, then other attributes take also specific values. Examples of DCs are: "a person who is pregnant is necessarily a woman" and "if it rains, then the road is certainly wet". This type of constraint may exist between features of feasible instances. Therefore, they may lead to dependencies between AXp's, and thus redundancies (as some follow from others).

In [8, 13], ICs (in the sense of constraints on possible feature vectors) were considered when generating abductive explanations while DCs (in the sense of dependencies between AXp's) were totally ignored. In this paper, we show that disregarding dependency constraints when explaining decisions may lead to exponentially more AXp's many of which may be redundant or superfluous. To bridge this gap, we investigate explanation functions that generate AXp's while taking into account both IC and DC constraints. Our contributions are fourfold: The first consists of proposing three novel types of abductive explanation that deal with constraints. They are generated from the whole set of instances that satisfy the constraints, thus discarding any instance that violates an integrity constraint. However, this is not sufficient for considering dependencies expressed by DCs. As a solution, the new types of explanation are based on the key notion of *coverage* of an explanation, i.e., the set of all instances it explains. Coverage is powerful enough to capture those constraints, and ensures the independence of explanations of every decision.

The second contribution consists of a thorough analysis of the complexity of explaining a decision. We show that finding a prime-implicant explanation becomes computationally much more challenging in a constrained setting.

The third contribution consists of proposing a paradigm for making the three solutions feasible. The idea is to avoid exhaustive search by examining a sample of the constrained feature space. We adapt the three types of explanations and show that the worst-case complexity of finding sample-based explanations is greatly reduced.

The fourth contribution consists of introducing desirable properties that an explanation function should satisfy, then comparing and analysing the novel functions against them. The results show, in particular, that when explanations are generated from a sample, complexity is greatly reduced but at the cost of violating a desirable

* Corresponding Author. Email: Martin.Cooper@irit.fr

Please contact the corresponding author for any appendices or supplementary material mentioned in the paper.

property, which ensures a kind of global coherence of the set of all explanations that may be returned by a function.

The paper is structured as follows: Section 2 recalls previous definitions of AXp and Section 3 discusses their limits. Section 4 defines the three novel types of AXp's, Section 5 analyses their complexity, and Section 6 studies their sample-based versions. Section 7 introduces properties of explainers and analyses the discussed functions against them. The last section concludes.

2 Background

Throughout the paper, we consider a *classification theory* as a tuple made of a finite set $\mathtt{F}$ of *features* (also called attributes), a function $\mathtt{dom}$ which returns the *domain* of every feature $f \in \mathtt{F}$, where $\mathtt{dom}(f)$ is finite and $|\mathtt{dom}(f)| > 1$, and a finite set $\mathtt{Cl}$ of *classes* with $|\mathtt{Cl}| \geq 2$. We call *literal* any pair (f, v) where $f \in \mathtt{F}$ and $v \in \mathtt{dom}(f)$. A *partial assignment* is any set of literals with each feature in $\mathtt{F}$ occurring at most once; it is called an *instance* when every feature appears once. We denote by $\mathbb{E}$ the set of all possible partial assignments and by $\mathbb{F}$ the *feature space*, i.e., the set of all instances. For all $E, E' \in \mathbb{E}$, the notation $E(E')$ is a shorthand for $E \subseteq E'$. The reason for this notation is that if E is a partial assignment, then E can be viewed as a predicate on instances x: $E(x)$ means that x agrees with E on the subset of features on which it is defined.

Definition 1 (Theory)**.** *A* classification theory *is a tuple* $\langle \mathtt{F}, \mathtt{dom}, \mathtt{Cl} \rangle$.

We consider a classifier κ, which is a function mapping every instance in $\mathbb{F}$ to a class in the set $\mathtt{Cl}$. We make the reasonable assumption that κ can be evaluated in polynomial time.

Abductive explanations (AXp) answer questions of the form: why is instance x assigned outcome c by classifier κ? They are partial assignments, which are sufficient for ensuring the prediction c. We recall below the definition of AXp [8, 25].

Definition 2 (wAXp, AXp)**.** *Let* $x \in \mathbb{F}$*. A* weak AXp *(wAXp) of* $\kappa(x)$ *is a partial assignment* $E \in \mathbb{E}$ *s.t.* $E(x)$ *and* $\forall y \in \mathbb{F}.(E(y) \rightarrow (\kappa(y) = \kappa(x)))$*. An* AXp *of* $\kappa(x)$ *is a subset-minimal weak AXp.*

Example 1. *Let* $\mathtt{F} = \{f_1, f_2\}$*, with* $\mathtt{dom}(f_1) = \mathtt{dom}(f_2) = \{0, 1\}$*, and* $\mathtt{Cl} = \{0, 1\}$*. Consider the classifier* κ_1 *s.t. for any* $x \in \mathbb{F}$*,* $\kappa_1(x) = (f_1, 1) \vee (f_2, 1)$*. Its predictions are given in the table below.*

	f_1	f_2	$\kappa_1(x_i)$
x_1	0	0	0
x_2	0	1	1
x_3	1	0	1
x_4	1	1	1

- $E_1 = \{(f_1, 1)\}$
- $E_2 = \{(f_2, 1)\}$
- $E_3 = \{(f_1, 1), (f_2, 1)\}$
- $E_4 = \{(f_1, 0), (f_2, 0)\}$

The decision $\kappa_1(x_4)$ *has three weak abductive explanations* E_1, E_2, E_3 *and two AXp's:* E_1, E_2*. The decision* $\kappa_1(x_1)$ *has a single wAXp/Axp, namely* E_4*.*

Explaining decisions made by classifiers is in general *not tractable* (assuming P≠NP) as shown in [8, 9].

Property 1. *([8, 9]) The problem of testing whether a set* $E \in \mathbb{E}$ *is a weak AXp is co-NP-complete. The problem of finding one AXp is in* FP^{NP}*, the class of functional problems that can be solved by a polynomial number of calls to a SAT oracle.*

In [8, 9, 13], the authors investigated *AXp's under constraints*. They assume, as we do in this paper, a finite set $\mathcal{C}$ of constraints between features, which can be considered as a predicate. For any partial assignment E, $\mathcal{C}(E)$ means that E satisfies all the constraints in $\mathcal{C}$, $\neg\mathcal{C}(E)$ means E violates at least one constraint, and if $\mathcal{C} = \emptyset$ then $\mathcal{C}(y) \equiv \top$. They took into account constraints in the definition of an AXp by checking only feasible instances which gave the following definition [8, 9, 13].

Definition 3 (AXpc)**.** *Let* $x \in \mathbb{F}$ *be s.t.* $\mathcal{C}(x)$*, where* $\mathcal{C}$ *is a finite set of constraints. A* weak AXpc *(wAXpc) of* $\kappa(x)$ *is a partial assignment* $E \in \mathbb{E}$ *such that:* $E(x)$*,* $\mathcal{C}(E)$*, and* $\forall y \in \mathbb{F}.(\mathcal{C}(y) \wedge E(y) \rightarrow (\kappa(y) = \kappa(x)))$*. An* AXpc *of* $\kappa(x)$ *is a subset-minimal weak AXpc.*

Example 1 (Cont) Assume the existence of the constraint $f_1 \wedge \neg f_2 \rightarrow \bot$, which means the instance x_3 is impossible. According to the above definitions, the decision $\kappa_1(x_1)$ has two weak AXpc's: E_4 and $E_5 = \{(f_2, 0)\}$, and one AXpc which is E_5.

The example shows an AXpc that is a subset of an AXp. The next result confirms this link between the two notions.

Proposition 1. *Let* $x \in \mathbb{F}$ *such that* $\mathcal{C}(x)$*, and* $E \in \mathbb{E}$*. If* E *is an AXp of* $\kappa(x)$*, then* $\exists E' \in \mathbb{E}$ *s.t.* $E' \subseteq E$ *and* E' *is an AXpc of* $\kappa(x)$*.*

However, an AXpc of a decision is not always related to an AXp of the same decision as shown below.

Example 2. *Consider the theory of Example 1 and the classifier* κ_2 *such that for* $x \in \mathbb{F}$*,* $\kappa_2(x) = \neg f_1$*. Suppose* $\mathcal{C}$ *contains one constraint,* $f_1 \wedge \neg f_2 \rightarrow \bot$*, which is violated by* x_3*.*

	f_1	f_2	$\kappa_2(x_i)$
x_1	0	0	1
x_2	0	1	1
x_3	1	0	0
x_4	1	1	0

- $E_1 = \{(f_1, 0)\}$
- $E_2 = \{(f_2, 0)\}$
- $E_3 = \{(f_1, 0), (f_2, 0)\}$

The decision $\kappa_2(x_1)$ *has* E_1 *as its sole AXp. However, it has two AXpc's:* E_1 *and* E_2*.*

3 Limits

The definition of AXp implicitly assumes *independence of features*, i.e. there are no constraints between the values they may take. The definition of an AXpc accounts for constraints but only partially. In what follows, we discuss three undesirable consequences of ignoring dependency constraints: existence of *superfluous* explanations, *redundancy* of explanations and *explosion in their number*.

Superfluous Explanations. We show next that ignoring constraints may lead to generating **gratuitous** explanations.

Example 2 (Cont) Recall that the decision $\kappa_2(x_1)$ has two AXpc's: $E_1 = \{(f_1, 0)\}$ and $E_2 = \{(f_2, 0)\}$. From the definition of κ_2 ($\forall x \in \mathbb{F}$, $\kappa_2(x) = \neg f_1$), it follows that E_1 is correct while E_2, although logically correct, is superfluous. The correlation between E_2 and $\kappa_2(x_1)$ is due to the dependency constraint stating: whenever f_2 takes the value 0, f_1 takes the same value (and consequently, κ_2 assigns 1 to the corresponding instance).

Redundancy. The following example shows that some AXp's may be *redundant* with respect to others due to dependency constraints between values of features.

Example 3. *Assume that* $\mathtt{F} = \{f_1, f_2\}$*, where* f_1 *and* f_2 *stand respectively for gender (0 for male, 1 for female) and being pregnant,* $\mathtt{dom}(f_1) = \mathtt{dom}(f_2) = \{0, 1\}$*, and* $\mathtt{Cl} = \{0, 1\}$*. Consider the constraint stating that only women can be pregnant. The instance* $\{(f_1, 0), (f_2, 1)\}$ *is then impossible. Consider the classifier* $\kappa_3(x) = f_1 \vee f_2$ *whose predictions for the possible instances are given in the table below.*

	f_1	f_2	$\kappa_3(x_i)$
x_1	0	0	0
x_2	1	0	1
x_3	1	1	1

- $E_1 = \{(f_1, 1)\}$
- $E_2 = \{(f_2, 1)\}$
- $E_3 = \{(f_1, 1), (f_2, 1)\}$

The decision $\kappa_3(x_3)$ *has two AXpc's:* E_1 *and* E_2*. Note that the two explanations are not independent, and* E_2 *is somehow redundant with* E_1 *since decisions concerning women in general hold for those who are pregnant.*

Exponential number of explanations. In the above examples, only one explanation is redundant. However, the number may be exponential as shown in the next example.

Example 4. *Let* $\mathtt{F} = \{f_1, \ldots, f_n\}$*,* $\forall i \in \{1, \ldots, n\}$*,* $\mathtt{dom}(f_i) = \{0, 1\}$*,* $\mathtt{Cl} = \{0, 1\}$*, and let* κ_4 *be the classifier* $\kappa_4(x) = f_n$*. Assume also that there is a constraint:* $f_n \equiv (\sum_{i=1}^{n-1} f_i \geq \lfloor n/2 \rfloor)$*. Let* $x = \{(f_i, 1) \mid i = 1, \ldots, n\}$*, so* $\kappa_4(x) = 1$*. The decision* $\kappa_4(x) = 1$ *has* $\binom{n}{k}$ *AXpc's, where* $k = \lfloor \frac{n}{2} \rfloor$*: all size-*$k$ *subsets of* $\{(f_i, 1) \mid i = 1, \ldots, n-1\}$*, as well as the AXpc* $\{(f_n, 1)\}$*. Observe that* $\{(f_n, 1)\}$ *subsumes all other AXpc's. Therefore, one could discard all explanations other than* $\{(f_n, 1)\}$ *since they are superfluous.*

To sum up, defining abductive explanations that deal with dependency constraints remains a challenge that has never been addressed in the literature. We propose in the next sections the first solutions and investigate their complexity and formal properties.

4 Explanations and feature-space coverage

We revisit in this section the definition of abductive explanations for constrained settings. In the rest of the paper, we assume a fixed but arbitrary classification theory $\langle \mathtt{F}, \mathtt{dom}, \mathtt{Cl} \rangle$ and a **finite** set $\mathcal{C}$ of constraints on the theory, and more precisely on its set $\mathbb{E}$ of partial assignments. For $E \in \mathbb{E}$, the notation $\mathcal{C}(E)$ means E satisfies all constraints in $\mathcal{C}$, $\neg\mathcal{C}(E)$ means E violates at least one constraint, and $\mathbb{F}[\mathcal{C}] = \{x \in \mathbb{F} \mid \mathcal{C}(x)\}$, i.e., the set of instances in $\mathbb{F}$ that satisfy the constraints. The set $\mathcal{C}$ satisfies the following properties:

(C1) $\mathbb{F}[\mathcal{C}] \neq \emptyset$ (constraints in $\mathcal{C}$ can be satisfied all together).
(C2) Let $E, E' \in \mathbb{E}$. If $E \subseteq E'$, then $\mathcal{C}(E') \rightarrow \mathcal{C}(E)$.

We consider a classifier κ which is a function mapping every instance in $\mathbb{F}[\mathcal{C}]$ to a class in $\mathtt{Cl}$. We assume that the test $x \in \mathbb{F}[\mathcal{C}]$ and the calculation of $\kappa(x)$ are polynomial.

We have seen that there are two types of constraints. Integrity constraints describe impossible assignments of values. The definition of an AXpc takes them into account by checking instance feasibility. Our approach starts by removing all unrealistic instances and focuses only on $\mathbb{F}[\mathcal{C}]$. However, we have seen in the previous section that this solution is not sufficient for dealing with dependencies between partial assignments that follow from constraints $\mathcal{C}$. Before showing how we deal with such dependency constraints, let us first define them.

Definition 4 (DC)**.** *A dependency constraint (DC) is any formula of the form* $E \rightarrow E'$ *such that:*

- $E, E' \in \mathbb{E} \setminus \{\emptyset\}$,
- $E \neq E'$,
- *For any* $x \in \mathbb{F}[\mathcal{C}]$*, if* $E(x)$ *then* $E'(x)$.

We denote by $\mathcal{C}^*$ *the set of all such constraints.*

DC's are defined on the entire set $\mathbb{F}[\mathcal{C}]$ of feasible instances. A DC $E \rightarrow E'$ means that whenever E holds, E' holds as well. In Example 3, the constraint $\{(f_2, 1)\} \rightarrow \{(f_1, 1)\}$ means that when the feature f_2 takes the value 1, the feature f_1 necessarily takes the same value.

Our approach takes advantage of such information for reducing the number of abductive explanations by avoiding dependent explanations, and therefore discarding redundant or superfluous ones. Before defining the novel notions of explanation, let us first introduce some useful notions. The first one is the *coverage* of a partial assignment, which is the set of instances it covers.

Definition 5 (Coverage)**.** *Let* $X \subseteq \mathbb{F}$ *and* $E \in \mathbb{E}$*. The* coverage *of* E *in* X *is the set* $\mathtt{cov}_X(E) = \{x \in X \mid E(x)\}$*. When* $X = \mathbb{F}[\mathcal{C}]$*, we write* $\mathtt{cov}(E)$ *for short.*

Example 3 (Cont) For $E = \{(f_1, 1)\}$, $\mathtt{cov}(E) = \{x_2, x_3\}$.

The second notion, which is crucial for the new definition of explanation, is a *subsumption* relation defined as follows.

Definition 6. *Let* $X \subseteq \mathbb{F}$ *and* $E, E' \in \mathbb{E}$*. We say that* E' subsumes E *in* X *if* $\forall x \in X.(E(x) \rightarrow E'(x))$*.* E' strictly subsumes E *in* X *if* E' *subsumes* E *in* X *but* E *does not subsume* E' *in* X*.*

We show that subsumption is closely related to coverage.

Proposition 2. *Let* $X \subseteq \mathbb{F}$ *and* $E, E' \in \mathbb{E}$*.*

- *The following statements are equivalent.*
  - E' *subsumes* E *in* X*.*
  - $\mathtt{cov}_X(E) \subseteq \mathtt{cov}_X(E')$.
- *If* $E' \subset E$*, then* E' *subsumes* E *in* X*. The converse does not always hold.*
- *If* $E \neq E'$*, then* $\mathtt{cov}_{\mathbb{F}}(E) \neq \mathtt{cov}_{\mathbb{F}}(E')$.

Example 3 (Cont) The partial assignment $E_1 = \{(f_1, 1)\}$ strictly subsumes $E_2 = \{(f_2, 1)\}$ in the space $\mathbb{F}[\mathcal{C}]$. Indeed, $\mathtt{cov}(E_1) = \{x_2, x_3\}$ and $\mathtt{cov}(E_2) = \{x_3\}$.

The subsumption relation is **not monotonic** meaning that a partial assignment E may subsume another (say E') in a set of instances X but not in some $Y \supset X$ as shown in the following example.

Example 1 (Cont) Assume again the existence of the constraint $f_1 \wedge \neg f_2 \rightarrow \bot$, which means $\mathbb{F}[\mathcal{C}] = \{x_1, x_2, x_4\}$. Let $E_1 = \{(f_1, 0)\}$ and $E_2 = \{(f_2, 0)\}$. Note that E_2 subsumes E_1 in $X = \{x_1\}$ but not in $\mathbb{F}[\mathcal{C}]$.

We show that every constraint in $\mathcal{C}^*$ can be expressed as a subsumption relation, which holds in any subset of instances.

Proposition 3. *Let $E, E' \in \mathbb{E}$. If $E \rightarrow E' \in \mathcal{C}^*$, then $\forall X \subseteq \mathbb{F}[\mathcal{C}]$, E' subsumes E in X.*

Let us now introduce our novel notion of *coverage-based prime-implicant explanation* (CPI-Xp). The first idea is to generate AXp's from the set of instances that satisfy the available constraints, thus discarding impossible instances. Furthermore, it selects AXp's which subsume the others, thus taking into account DCs. This is equivalent to selecting AXp's that apply to more instances in $\mathbb{F}[\mathcal{C}]$.

Definition 7 (CPI-Xp)**.** *Let $x \in \mathbb{F}[\mathcal{C}]$. A* coverage-based PI-explanation *(CPI-Xp) of $\kappa(x)$ is any $E \in \mathbb{E}$ such that:*

- $E(x)$,
- $\forall y \in \mathbb{F}[\mathcal{C}].(E(y) \rightarrow (\kappa(y) = \kappa(x)))$,
- $\nexists E' \in \mathbb{E}$ *such that E' satisfies the above conditions and strictly subsumes E in $\mathbb{F}[\mathcal{C}]$.*

While a CPI-Xp is clearly a weak AXpc, the two notions do not always coincide when the set of constraints is empty.

Example 1 (Cont) Let $\mathcal{C} = \emptyset$ ($\mathbb{F}[\mathcal{C}] = \mathbb{F}$). Note that $\texttt{cov}(E_1) = \{x_3, x_4\}$, $\texttt{cov}(E_2) = \{x_2, x_4\}$, $\texttt{cov}(E_3) = \{x_4\}$ showing that E_3 is strictly subsumed by E_1 and E_2 in $\mathbb{F}$. So E_3 is not a CPI-Xp of $\kappa_1(x_4)$ while it is a wAXpc and a wAXp.

Let us now show how the notion of CPI-Xp solves the three problems discussed in the previous section.

Example 2 (Cont) Recall that $f_1 \wedge \neg f_2 \rightarrow \bot$ is a constraint, so $\{(f_1, 1)\} \rightarrow \{(f_2, 1)\} \in \mathcal{C}^*$ and $\mathbb{F}[\mathcal{C}] = \{x_1, x_2, x_4\}$. The decision $\kappa_2(x_1)$ has three weak AXpc's in $\mathbb{F}[\mathcal{C}]$: $E_1 = \{(f_1, 0)\}$, $E_2 = \{(f_2, 0)\}$, and $E_3 = \{(f_1, 0), (f_2, 0)\}$. Note that $\texttt{cov}(E_2) = \texttt{cov}(E_3) = \{x_1\} \subset \texttt{cov}(E_1) = \{x_1, x_2\}$. So E_1 is the sole CPI-Xp of $\kappa_2(x_1)$, discarding the superfluous AXpc E_2. This shows that subsumption is powerful enough to detect **gratuitous correlations** between features and decisions.

Example 3 (Cont) Recall that $\mathcal{C}^* = \{E_2 \rightarrow E_1\}$, $E_1 = \{(f_1, 1)\}$, $E_2 = \{(f_2, 1)\}$, and $\mathbb{F}[\mathcal{C}] = \{x_1, x_2, x_3\}$. The decision $\kappa_3(x_3)$ has two AXpc's: E_1 and E_2. However, it has a single coverage-based PI-explanation, namely E_1 which subsumes E_2. The redundant AXpc E_2 is thus discarded.

Example 4 (Cont) Recall that $x = \{(f_i, 1) \mid 1 \leq i \leq n\}$ and the decision $\kappa_4(x)$=1 has a combinatorial number of AXpc's: all subsets of $\{(f_i, 1) \mid 1 \leq i \leq n-1\}$ of size $\binom{n}{\lfloor n/2 \rfloor}$. Due to the constraint $f_n \equiv (\sum_{i=1}^{n-1} f_i \geq \lfloor n/2 \rfloor)$, the decision $\kappa_4(x)$=1 has a single CPI-Xp, namely $\{(f_n, 1)\}$. So, there is a **drastic reduction** in the number of explanations.

Despite a significant reduction in the number of abductive explanations (AXp's), a decision may still have several coverage-based PI-explanations. In what follows, we discuss two **criteria** to further reduce the number of CPI-Xps: *conciseness* and *generality*. Let us start with the first criterion. We show that CPI-Xp may contain irrelevant information.

Example 5. *Consider a theory made of three binary features f_1, f_2, f_3. Let $E_1 = \{(f_1, 1), (f_2, 1)\}$ and $E_2 = \{(f_3, 1)\}$. Assume $\mathcal{C}^* = \{E_1 \rightarrow E_2, E_2 \rightarrow E_1\}$, then $\texttt{cov}(E_1) = \texttt{cov}(E_2) = \texttt{cov}(E_1 \cup E_2)$. Suppose that E_2 is a CPI-Xp. Then $E_1 \cup E_2$ is also a CPI-Xp but is not subset-minimal.*

Concision of explanations is important given the well known cognitive limitations of human users when processing information [20]. A common way for ensuring concision is to require *minimality* in order to avoid irrelevant information in an explanation.

Definition 8 (Minimal CPI-Xp)**.** *Let $x \in \mathbb{F}[\mathcal{C}]$. A* minimal coverage-based PI-explanation *(mCPI-Xp) of $\kappa(x)$ is a subset-minimal CPI-Xp of $\kappa(x)$.*

Example 5 (Cont) The set $E_1 \cup E_2$ is not a minimal CPI-Xp since E_1 and E_2 are CPI-Xps.

Let us now turn our attention to the generality criterion, which concerns the coverage of an explanation. Example 5 shows that coverage-based PI-explanations of a decision may have exactly the same coverage. Indeed, if E_1 is a minimal CPI-Xp, then so is E_2, and both have the same coverage. We say that such explanations are *equivalent*.

Definition 9 (Equivalence)**.** *Let $X \subseteq \mathbb{F}[\mathcal{C}]$. Two sets $E, E' \in \mathbb{E}$ are* equivalent *in X, denoted by $E \approx E'$, iff they subsume each other in the set X.*

Notation: Let X be a set and $\approx$ an equivalence relation on X. A set of *representatives* of X is a subset of X containing exactly one element of every equivalence class of X, i.e., one element among equivalent ones.

We propose next *preferred coverage-based explanations* that consider only one mCPI-Xp among equivalent ones (obviously in $\mathbb{F}[\mathcal{C}]$ since mCPI-Xp are produced from $\mathbb{F}[\mathcal{C}]$).

Definition 10 (Preferred CPI-Xp)**.** *Let $x \in \mathbb{F}[\mathcal{C}]$. A* preferred coverage-based PI-explanation *(pCPI-Xp) of $\kappa(x)$ is a representative of the set of miminal CPI-Xp's of $\kappa(x)$.*

Example 5 (Cont) Definition 10 selects either E_1 or E_2 (but not both) as a pCPI-Xp.

The three novel notions of explanation are clearly related to each other. Furthermore, when the set of constraints is empty, AXpc explanations presented in [8] (see Def. 3) coincide with both AXp's and minimal CPI-Xp's. In the general case, a mCPI-Xp is an AXpc but the converse does not hold. This confirms that AXpc deals only with integrity constraints and ignores dependency constraints. Before presenting a summary of the links, let us first introduce the notion of *explanation function* or *explainer*.

Definition 11 (Explainer)**.** *An explainer is a function $\mathbf{L}_y$ mapping every instance $x \in \mathbb{F}[\mathcal{C}]$ into the subset of $\mathbb{E}$ consisting of y-explanations of the decision $\kappa(x)$, for $y \in \{wAXp, AXp, wAXpc, AXpc, CPI\text{-}Xp, mCPI\text{-}Xp, pCPI\text{-}Xp\}$.*

Proposition 4. *Let $x \in \mathbb{F}[\mathcal{C}]$.*

1. $\mathbf{L}_{AXp}(x) \subseteq \mathbf{L}_{wAXp}(x)$,
2. $\mathbf{L}_{CPI-Xp}(x) \subseteq \mathbf{L}_{wAXpc}(x)$,
3. $\mathbf{L}_{mCPI-Xp}(x) \subseteq \mathbf{L}_{AXpc}(x) \subseteq \mathbf{L}_{wAXpc}(x)$,
4. $\mathbf{L}_{pCPI-Xp}(x) \subseteq \mathbf{L}_{mCPI-Xp}(x) \subseteq \mathbf{L}_{CPI-Xp}(x)$,
5. *If $\mathcal{C} = \emptyset$, then $\mathbf{L}_{AXp}(x) = \mathbf{L}_{AXpc}(x) = \mathbf{L}_{mCPI-Xp}(x)$.*

To sum up, we introduced three novel types of abductive explanations that better take into account constraints, and solve the three problems (superfluous, redundant, exponential number of explanations) of the existing definitions.

5 Complexity analysis

Let us investigate the computational complexity of the new types of explanation. We focus on the complexity of *testing* whether a given partial assignment is a (minimal, preferred) coverage-based PI-explanation, and the complexity of *finding* one such explanation.

We first consider the computational problem of deciding if a partial assignment is a coverage-based PI-explanation (CPI-Xp). We show that the problem can be rewritten as an instance of $\forall\exists$SAT, the problem of testing the satisfiability of a quantified boolean formula of the form $\forall x \exists y \phi(x, y)$, where x, y are vectors of boolean variables and ϕ is an arbitrary boolean formula with no free variables other than x and y. It is well known that $\forall\exists$SAT is complete for the complexity class Π_2^{P}. It turns out that testing whether a weak AXpc is a coverage-based PI-explanation is also Π_2^{P}-complete.

Theorem 1. *The problem of testing whether a weak AXpc E is a coverage-based PI-explanation is Π_2^{P}-complete.*

We now consider the problem of actually finding a coverage-based PI-explanation. In the supplementary material we give an algorithm which returns one CPI-Xp of a given decision $\kappa(v) = c$. It is based on the following idea: if a weak AXpc E is not a coverage-based PI-explanation, then this is because there is a weak AXpc E' that strictly subsumes E. We call such an E' a counter-example to the hypothesis that E is a coverage-based PI-explanation. Therefore, starting from a weak AXpc E, we can look for a counter-example E': if no counter-example exists then we return E, otherwise we can replace E by E' and re-iterate the process. This loop must necessarily halt since there cannot be an infinite sequence of partial assignments $E_1, E_2, \ldots$ such that E_{i+1} strictly subsumes E_i $(i = 1, 2, \ldots)$. We can be more specific: the following proposition shows that the number of iterations is bounded by the number of features.

Theorem 2. *Let $n = |\mathtt{F}|$. A CPI-Xp can be found by n calls to an oracle for testing whether a given weak AXpc is a coverage-based PI-explanation.*

It follows that the complexity of finding one coverage-based PI-explanation is essentially the same (modulo a linear factor) as testing whether a given weak AXpc is a coverage-based PI-explanation.

The following proposition shows that imposing minimality (for set inclusion) does not change the complexity. A minimal CPI-Xp can be found by n calls to an oracle (for testing whether a given weak AXpc is a CPI-Xp) together with $2n$ calls to a SAT oracle. Note that since finding a preferred CPI-Xp (i.e. pCPI-Xp) consists of finding one minimal CPI-Xp, there is no change in complexity. The difference between pCPI-Xp's and minimal CPI-Xp's becomes apparent when enumerating all explanations: there can be many less pCPI-Xp's which is an advantage for the user.

Theorem 3. *Let $n = |\mathtt{F}|$. A mCPI-Xp (resp. pCPI-Xp) can be found by n calls to an oracle for testing whether a given weak AXpc is a coverage-based PI-explanation together with $2n$ calls to a SAT oracle.*

To sum up, we have shown that taking into account constraints (in the definition of coverage-based prime implicants) may produce less-redundant explanations, but at the cost of a potential increase in computational complexity.

6 Sample-based explanations

When a classifier is a black-box or a deep neural network which cannot be realistically written down as a function, the only algorithm for testing whether a set of literals is a (weak) AXp is an exhaustive search over the whole feature space. This explains why they are costly from a computational point of view. We have seen in the previous section that the computational complexity of the three novel types of explanations that deal with constraints (CPI-Xp, mCPI-Xp and pCPI-Xp) is even worse. In this section, we propose a paradigm for making the solutions feasible. The idea is to avoid the exhaustive search by examining only a sample of the feature space. The obtained explanations are approximations that can be obtained with lower complexity as we will see next.

In this section, we concentrate on a sample (or dataset) $\mathcal{T} \subseteq \mathbb{F}[\mathcal{C}]$ and the associated values of a black-box classifier κ. Note that $\mathcal{T}$ may be the dataset a classifier has been trained on, a dataset on which the classifier has better performance, or may be generated in a specific way as in [23, 24]. However, we assume that every possible class in $\mathtt{Cl}$ is considered in the sample, i.e., it is assigned to at least one instance in $\mathcal{T}$: $\forall c \in \mathtt{Cl}, \exists x \in \mathcal{T}$ such that $\kappa(x) = c$.

In what follows, we adapt the definitions of the different explanations discussed in the previous sections, and add a suffix 'd-' to indicate the new versions.

6.1 Abductive explanations based on samples

Recall that an AXp E of a decision $\kappa(x) = c$ is a minimal subset of x which guarantees the class c. If κ is a black-box function, then testing this definition for a given E requires testing the exponential number of assignments to the features not assigned by the partial assignment E. The following definition only requires us to test those instances in the sample $\mathcal{T}$.

Definition 12 (d-wAXp, d-AXp)**.** *Let $x \in \mathbb{F}$. A* weak dataset-based AXp *(d-wAXp) of $\kappa(x)$ wrt to $\mathcal{T}$ is a partial assignment $E \in \mathbb{E}$ such that $E(x)$ and $\forall y \in \mathcal{T}$, if $E(y)$ then $\kappa(y) = \kappa(x)$. A* dataset-based AXp *(d-AXp) of $\kappa(x)$ is a subset-minimal d-wAXP of $\kappa(x)$.*

In other words, a d-AXp is mathematically equivalent to an AXp under the artificial constraint that the only allowed feature vectors are those in the dataset.

Example 6. *Assume a theory made of four binary features, a binary classifier κ defined as follows: $\kappa(x) = (f_1 \wedge f_2) \vee (f_3 \wedge f_4)$. Let $\mathcal{T}$ be a sample whose instances and their predictions are summarized in the table below.*

$\mathcal{T}$	f_1	f_2	f_3	f_4	κ
x_1	0	1	0	0	0
x_2	0	1	0	1	0
x_3	0	1	1	0	0
x_4	0	0	1	1	1
x_5	1	1	1	1	1
x_6	1	1	0	1	1

- $E_1 = \{(f_1, 1), (f_2, 1)\}$
- $E_2 = \{(f_3, 1), (f_4, 1)\}$
- $E_3 = \{(f_1, 1)\}$

The decision $\kappa(x_5) = 1$ has two AXp's (over the whole feature space $\mathbb{F}$): E_1 and E_2. It has two d-AXp's: E_2 and E_3. The fact that E_3 is a d-AXp is a consequence of the fact that the pair $(f_1, 1)$, $(f_2, 0)$ never occurs in instances of $\mathcal{T}$. Note that E_3 is a d-wAXP while it is not a weak AXp.

We have seen in Property 1 that the problems of testing and finding abductive explanations are not tractable. We show next that

their **sample-based versions are tractable**. Indeed, there is an obvious algorithm (by applying directly the definition) with complexity $O(mn)$ for testing whether a set of literals E is a weak dataset-based AXp (d-wAXp), where n and m stand for the number of features and instances in the dataset $\mathcal{T}$ respectively. We can also test whether a weak d-AXp E is subset-minimal in $O(mn^2)$ time by testing if E remains a weak d-AXp after deletion of each literal. Indeed, as for AXp's [8], a d-AXp can be found by starting with $E = x$, the instance to be explained, and in turn for each of the n elements of E, delete it if E remains a weak d-AXp after its deletion. It follows that a d-AXp can be found in $O(mn^2)$ time.

Theorem 4. *Let $n = |\mathbf{F}|$, $m = |\mathcal{T}|$ and $E \in \mathbb{E}$.*

- *Testing whether E is a d-wAXp can be achieved in $O(mn)$ time.*
- *Finding a d-AXp can be achieved in $O(mn^2)$ time.*

6.2 Coverage-based explanations based on samples

We now study the sample-based versions of the three types of explanations that take into account the coverage of explanations. Coverage is now among instances in the dataset $\mathcal{T} \subseteq \mathbb{F}[\mathcal{C}]$.

Definition 13 (d-CPI-Xp, d-mCPI-Xp, d-pCPI-Xp)**.** *Let $x \in \mathbb{F}[\mathcal{C}]$. A* dataset-based CPI-explanation *(d-CPI-Xp) of $\kappa(x)$ is a partial assignment $E \in \mathbb{E}$ such that:*

- $E(x)$,
- $\forall y \in \mathcal{T}.(E(y) \rightarrow (\kappa(y) = \kappa(x)))$,
- $\nexists E' \in \mathbb{E}$ *such that E' satisfies the above conditions and strictly subsumes E in $\mathcal{T}$.*

A dataset-based minimal CPI-explanation *(d-mCPI-Xp) of $\kappa(x)$ is a subset minimal d-CPI-Xp of $\kappa(x)$. A* dataset-based preferred CPI-explanation *(d-pCPI-Xp) of $\kappa(x)$ is a representative of the set of d-mCPI-Xp's of $\kappa(x)$ in $\mathcal{T}$.*

Proposition 5. *The following inclusions hold:*

- $\mathbf{L}_{d-CPI-Xp}(x) \subseteq \mathbf{L}_{d-wAXpc}(x)$
- $\mathbf{L}_{d-mCPI-Xp}(x) \subseteq \mathbf{L}_{d-AXpc}(x)$
- $\mathbf{L}_{d-pCPI-Xp}(x) \subseteq \mathbf{L}_{d-mCPI-Xp}(x) \subseteq \mathbf{L}_{d-CPI-Xp}(x)$

Even when the set of constraints $\mathcal{C}$ is empty, dataset-based AXp's do not coincide with dataset-based CPI-Xp's or mCPI-Xp's. This is mainly due to the notion of subsumption which privileges explanations with greater coverage.

Example 2 (Cont) Consider again the theory below and recall that for $x \in \mathbb{F}$, $\kappa_2(x) = \neg f_1$. Suppose $\mathcal{C} = \emptyset$ and let us focus on the sample $\mathcal{T} = \{x_1, x_2, x_3\}$ (x_4 being discarded).

	f_1	f_2	$\kappa_2(x_i)$
x_1	0	0	1
x_2	0	1	1
x_3	1	0	0
x_4	1	1	0

- $E_1 = \{(f_1, 0)\}$
- $E_2 = \{(f_2, 1)\}$
- $E_3 = \{(f_1, 0), (f_2, 1)\}$

The decision $\kappa_2(x_2)$ has three d-wAXps (E_1, E_2, E_3) and two d-AXp's (E_1, E_2). However, it has a single d-CPI-Xp/d-mCPI-Xp: E_1. Indeed, its coverage in $\mathcal{T}$ is $\{x_1, x_2\}$, which is a super-set of the coverage $\{x_2\}$ of E_2, E_3.

We show that considering coverage in the definition of prime implicant does not greatly increase the complexity of finding explanations based on the dataset. There is a polynomial-time algorithm for testing whether a partial assignment E is a d-CPI-Xp and indeed for finding a d-CPI-Xp. Furthermore, finding a subset-minimal d-CPI-Xp is asymptotically no more costly than finding a d-CPI-Xp.

Theorem 5. *Let $E \in \mathbb{E}$.*

- *Testing whether E is a d-CPI-Xp can be achieved in $O(m^2 n)$ time.*
- *Finding a d-CPI-Xp, a minimal d-CPI-Xp and a preferred d-CPI-Xp can be achieved in $O(m^2 n^2)$ time.*

Table 1 summarizes all the complexity results concerning the different types of AXps' reviewed so far.

Explanation	Complexity of testing	Complexity of finding one
d-wAXp	P	polytime
d-AXp	P	polytime
d-CPI-Xp	P	polytime
d-mCPI-Xp	P	polytime
d-pCPI-Xp	P	polytime
wAXp	co-NP-complete	polytime
AXp	P^{NP}	FP^{NP}
CPI-Xp	Π_2^P-complete	$\text{FP}^{\Sigma_2^{\text{p}}}$
mCPI-Xp	Π_2^P-complete	$\text{FP}^{\Sigma_2^{\text{p}}}$
pCPI-Xp	Π_2^P-complete	$\text{FP}^{\Sigma_2^{\text{p}}}$

Table 1. Complexities of testing/finding different explanations. $\text{FP}^{\mathcal{L}}$ is the class of function problems that can be solved by a polynomial number of calls to an oracle for the language $\mathcal{L}$. We assume a white box, i.e. κ is an arbitrary but *known* function, except for the case of sample-based explanations (where κ may be a black-box function).

To sum up, the previous definitions and results show that the sample-based approach presents three advantages: i) testing the validity of an explanation is linear in the size of the sample whatever the function κ and the constraints, ii) it can be applied even when the classifier is a black-box, iii) sample-based abductive explanations may be smaller and hence easier to interpret for a human user.

7 Properties of explanation functions

We have seen that for each type of explanation studied in this paper, there is a corresponding explanation function $\mathbf{L}$ mapping every instance (in $\mathbb{F}$ or $\mathbb{F}[\mathcal{C}]$) into a subset of $\mathbb{E}$. In this section, we provide seven desirable properties for an explanation function. The first four properties have counterparts in [2], where explanation functions explain the *global* behaviour of a classifier in a non-constrained setting, and so answer the question: "why does a classifier recommend a given class in general?" We adapt the properties for explaining individual decisions and introduce three novel ones that concern how a function should deal with constraints.

Definition 14. *Let $\mathbf{L}$ be an explanation function.*

(Success) $\forall x \in \mathbb{F}[\mathcal{C}]$, $\mathbf{L}(x) \neq \emptyset$.
(Non-Triviality) $\forall x \in \mathbb{F}[\mathcal{C}]$, $\forall E \in \mathbf{L}(x)$, $E \neq \emptyset$.
(Irreducibility) $\forall x \in \mathbb{F}[\mathcal{C}]$, $\forall E \in \mathbf{L}(x)$, $\forall l \in E$, $\exists x' \in \mathbb{F}[\mathcal{C}]$ *such that* $\kappa(x') \neq \kappa(x)$ *and* $(E \setminus \{l\})(x')$.
(Coherence) $\forall x, x' \in \mathbb{F}[\mathcal{C}]$ *such that* $\kappa(x) \neq \kappa(x')$, $\forall E \in \mathbf{L}(x)$, $\forall E' \in \mathbf{L}(x')$, $\nexists x'' \in \mathbb{F}[\mathcal{C}]$ *s.t.* $(E \cup E')(x'')$.

	wAXp	AXp	AXpc	CPI-Xp	mCPI-Xp	pCPI-Xp	dCPI-Xp	dmCPI-Xp	dpCPI-Xp	d-wAXp	d-AXp
Success	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Non-Triv.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Irreduc.	×	✓	✓	×	✓	✓	×	✓	✓	×	✓
Coherence	✓	✓	✓	✓	✓	✓	×	×	×	×	×
Consist.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Indep.	×	×	×	✓	✓	✓	×	×	✓	×	×
Non-Equiv.	✓	✓	×	×	×	✓	×	×	✓	×	×

Table 2. The symbol ✓ (resp. ×) stands for satisfied (resp. violated).

(Consistency) $\forall x \in \mathbb{F}[\mathcal{C}], \forall E \in \mathbf{L}(x), \mathcal{C}(E)$ *holds.*
(Independence) $\forall x \in \mathbb{F}[\mathcal{C}], \nexists E, E' \in \mathbf{L}(x)$ *such that* $E \rightarrow E' \in \mathcal{C}^*$ *and* $E' \rightarrow E \notin \mathcal{C}^*$.
(Non-Equivalence) $\forall x \in \mathbb{F}[\mathcal{C}], \forall E, E' \in \mathbf{L}(x), E \not\approx E'$.

Success ensures existence of explanations. *Non-Triviality* discards empty explanations as they are non-informative. *Irreducibility* states that an explanation should not contain unnecessary information. *Coherence* ensures a global compatibility of the explanations provided for all the instances. It avoids erroneous explanations. Consider a function κ which classifies animals as mammals or not, where animals are described by n features such as *milks its young*, *lays eggs*, etc. Let x be a mouse and x' an eagle. If the explanation E for $\kappa(x) = 1$ is that mice milk their young and the explanation E' for $\kappa(x') = 0$ is that eagles lay eggs, then coherence is not satisfied because there are animals x'' (such as the platypus) which milk their young and lay eggs. *Consistency* ensures that explanations satisfy all constraints in $\mathcal{C}$. *Independence* ensures that dependency constraints are considered and the explanations of a decision should be pairwise independent. *Non-equivalence* avoids equivalent explanations. This property is important for reducing the number of explanations.

We show that the seven properties are *compatible*, i.e., they can be satisfied all together by an explanation function.

Theorem 6. *The properties are compatible.*

Table 2 summarizes the properties satisfied by each type of explanation discussed in the paper. It thus provides a comprehensive formal comparison of their explainers, and sheds light on the key properties that distinguish any pair of explainers.

Theorem 7. *The properties of Table 2 hold.*

The results confirm that existing definitions of abductive explanations ignore dependency constraints (they violate independence). However, they satisfy in a vacuous way consistency since the latter deals only with feasible instances (elements of $\mathbb{F}[\mathcal{C}]$). The three novel types that are generated from the feature space handle properly integrity constraints (satisfaction of consistency) and dependency constraints (satisfaction of independence).

The results show also that among the new types, only the two versions of preferred CPI-Xp satisfy non-equivalence. Hence, they use the most discriminatory selection criterion. Non-equivalence is surprisingly satisfied by wAXp and AXp because they consider the entire feature space, and in this particular case, two different partial assignments can never have the same coverage. Note that the property is violated by their sample-based versions.

Another result which is due to the use of the whole feature space concerns the satisfaction of coherence. The property is lost when explanations are based on a dataset, thus erroneous explanations may be provided for decisions. The main reason behind this issue is that we generate explanations under incomplete information, and thus some explanations may not hold when tested on the whole feature space. Another consequence of incompleteness of information is that the sample-based versions of CPI-Xp and mCPI-Xp violate independence due to missing instances in the sample.

To sump up, the explainer that generates preferred CPI-Xp is the only one that satisfies all the properties.

8 Related work

As explained in the introduction, abductive explanations have largely been studied in the XAI literature. However, to the best of our knowledge only a few works ([8, 13]) have considered the constrained setting. We have shown that those works deal only partially with constraints as they ignore dependency constraints.

Many papers [7, 12, 17, 18] are concerned with explaining the inconsistency of constraints, which is quite far from the problem we are studying (explaining the output of a classifier κ in a constrained feature-space). [16] add new constraints in a SAT solver in order to quickly find explanations of a classifier, again a different problem.

9 Conclusion

Our work is the first to wholly take into account constraints for producing abductive explanations. A general conclusion that can be drawn from our study is that constraints may lead to less-redundant explanations, but at the cost of a potential increase in complexity. Another conclusion is that sample-based versions of explanations provide a tractable alternative, especially in the case of black-box classifiers. The downside of the approach is, unfortunately, explanations are only valid for the instances in the dataset and not for the whole feature space. Therefore, there is a trade-off between computational complexity and coherence of explanations.

This work can be extended in different ways. We plan to characterize the whole family of explainers that satisfy all (or a subset of) the properties. We also plan to define sample-based explainers that consider constraints and satisfy the coherence property. Another avenue of future research is to learn constraints from the dataset: we would limit ourselves to small-arity constraints to make this feasible.

Recall that the set of all pCPI-Xp's contains a single representative from each equivalence class, where two explanations are equivalent if they cover the same set of instances. A challenging open problem is the enumeration of pCPI-Xp's, which corresponds to enumerating abductive explanations whose coverages are all pairwise incomparable for subset inclusion.

Acknowledgments

This work was supported by the AI Interdisciplinary Institute ANITI, funded by the French program "Investing for the Future – PIA3" under grant agreement no. ANR-19-PI3A-0004.

References

[1] Leila Amgoud, 'Explaining black-box classifiers: Properties and functions', *International Journal of Approximate Reasoning*, **155**, 40–65, (2023).

[2] Leila Amgoud and Jonathan Ben-Naim, 'Axiomatic foundations of explainability', in *IJCAI*, pp. 636–642, (2022).

[3] Leila Amgoud, Philippe Muller, and Henri Trenquier, 'Leveraging argumentation for generating robust sample-based explanations', in *IJCAI*, p. In press, (2023).

[4] Gilles Audemard, Steve Bellart, Louenas Bounia, Frederic Koriche, Jean-Marie Lagniez, and Pierre Marquis, 'On preferred abductive explanations for decision trees and random forests', in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pp. 643–650, (2022).

[5] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis, 'Trading complexity for sparsity in random forest explanations', in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pp. 5461–5469. AAAI Press, (2022).

[6] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski, 'Computing abductive explanations for boosted trees', in *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4699–4711, (2023).

[7] James Bailey and Peter J. Stuckey, 'Discovery of minimal unsatisfiable subsets of constraints using hitting set dualization', in *Practical Aspects of Declarative Languages, 7th International Symposium, PADL*, eds., Manuel V. Hermenegildo and Daniel Cabeza, volume 3350, pp. 174–186. Springer, (2005).

[8] Martin C. Cooper and João Marques-Silva, 'On the tractability of explaining decisions of classifiers', in *CP 2021*, ed., Laurent D. Michel, volume 210 of *LIPIcs*, pp. 21:1–21:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, (2021).

[9] Martin C. Cooper and João Marques-Silva, 'Tractability of explaining classifier decisions', *Artificial Intelligence*, **316**, 103841, (2023).

[10] Adnan Darwiche and Auguste Hirth, 'On the reasons behind decisions', in *24th European Conference on Artificial Intelligence ECAI*, volume 325, pp. 712–720. IOS Press, (2020).

[11] Alexis de Colnet and Pierre Marquis, 'On the complexity of enumerating prime implicants from decision-dnnf circuits', in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, ed., Luc De Raedt, pp. 2583–2590, (2022).

[12] Maria J. García de la Banda, Peter J. Stuckey, and Jeremy Wazny, 'Finding all minimal unsatisfiable subsets', in *Proceedings of the 5th International ACM SIGPLAN Conference on Principles and Practice of Declarative Programming*, pp. 32–43. ACM, (2003).

[13] Niku Gorji and Sasha Rubin, 'Sufficient reasons for classifier decisions in the presence of domain constraints', in *AAAI*, pp. 5660–5667, (2022).

[14] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva, 'Abduction-based explanations for machine learning models', in *AAAI 2019*, pp. 1511–1519. AAAI Press, (2019).

[15] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva, 'On relating explanations and adversarial examples', in *Thirty-third Conference on Neural Information Processing Systems, NeurIPS*, pp. 15857–15867, (2019).

[16] Alexey Ignatiev and João P. Marques Silva, 'Sat-based rigorous explanations for decision lists', in *Theory and Applications of Satisfiability Testing - SAT*, eds., Chu-Min Li and Felip Manyà, volume 12831, pp. 251–269. Springer, (2021).

[17] Ulrich Junker, 'QUICKXPLAIN: preferred explanations and relaxations for over-constrained problems', in *Proceedings of the Nineteenth National Conference on Artificial Intelligence AAAI*, eds., Deborah L. McGuinness and George Ferguson, pp. 167–172, (2004).

[18] Mark H. Liffiton and Karem A. Sakallah, 'On finding all minimally unsatisfiable subformulas', in *Theory and Applications of Satisfiability Testing, 8th International Conference, SAT*, eds., Fahiem Bacchus and Toby Walsh, volume 3569 of *Lecture Notes in Computer Science*, pp. 173–186. Springer, (2005).

[19] Xinghan Liu and Emiliano Lorini, 'A unified logical framework for explanations in classifier systems', *Journal of Logic and Computation*, **33**(2), 485–515, (2023).

[20] George A. Miller, 'Some limits on our capacity for processing information', *Psychological review*, **63**(2), 81–97, (1956).

[21] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, **267**, 1–38, (2019).

[22] C. Molnar, *Interpretable Machine Learning*, Lulu.com, 2020.

[23] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Why should I trust you?: Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135–1144, (2016).

[24] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Anchors: High-precision model-agnostic explanations', in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 1527–1535, (2018).

[25] Andy Shih, Arthur Choi, and Adnan Darwiche, 'A symbolic approach to explaining Bayesian network classifiers', in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pp. 5103–5111, (2018).

[26] Bernhard Thalheim, *Dependencies in Relational Databases*, Vieweg+Teubner Verlag Wiesbaden, 1991.