



**HAL**  
open science

# Axiomatic Characterisations of Sample-based Explainers

Leila Amgoud, Martin Cooper, Debbaoui Salim

► **To cite this version:**

Leila Amgoud, Martin Cooper, Debbaoui Salim. Axiomatic Characterisations of Sample-based Explainers. 27TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, ECAI-24, Oct 2024, Santiago de Compostela, Spain. <hal-04729098>

**HAL Id: hal-04729098**

**<https://hal.science/hal-04729098v1>**

Submitted on 9 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Axiomatic Characterisations of Sample-based Explainers

Leila Amgoud<sup>a</sup>, Martin Cooper<sup>b,\*</sup> and Salim Debbaoui<sup>b</sup>

<sup>a</sup>IRIT, CNRS, France

<sup>b</sup>IRIT, Toulouse University, France

**Abstract.** Explaining decisions of black-box classifiers is both important and computationally challenging. In this paper, we scrutinize explainers that generate *feature-based explanations* from samples or datasets. We start by presenting a set of desirable properties that explainers would ideally satisfy, delve into their relationships, and highlight incompatibilities of some of them. We identify the entire family of explainers that satisfy two key properties which are compatible with all the others. Its instances provide sufficient reasons, called *weak abductive explanations*. We then unravel its various sub-families that satisfy subsets of compatible properties. Indeed, we fully characterize all the explainers that satisfy any subset of compatible properties. In particular, we introduce the first (broad family of) explainers that guarantee the existence of explanations and their global consistency. We discuss some of its instances including the *irrefutable explainer* and the *surrogate explainer* whose explanations can be found in polynomial time.

## 1 Introduction

In recent years, AI systems have demonstrated remarkable capabilities but this success has often come at the expense of a corresponding lack of explainability. Indeed, deep neural networks behave effectively like black-box functions, meaning that explaining their decisions is inherently intractable. On the other hand, legislators have recognised the importance of providing explanations to end-users affected by decisions taken by AI systems [20, 37]. Thus, the question is how to reconcile the user’s right to an explanation with the intractability of the corresponding computational problem.

In this paper, we concentrate on decisions taken by classifiers learnt by Machine-Learning (ML). There have been some important successes in identifying ML models which allow either polynomial-time explainability [5, 12, 16, 27, 30] or effectively-tractable explainability via efficient solvers [8, 9, 23, 26]. However, success of such formally-correct explaining on neural networks has been limited to relatively small networks [25]. One solution to the right-to-an-explanation/intractability-of-explaining dilemma is the use of *surrogate models*: instead of explaining the decision of classifier  $\kappa$  on an instance  $x$ , the back-box function  $\kappa$  is approximated by a simpler explainable function  $\tilde{\kappa}$  in the neighbourhood of  $x$  [33, 34].

Another solution is to produce explanations which are valid on a subset of feature space [1, 3, 14]. The argument behind this endeavour is that since ML models are obtained from a dataset rather than a complete search of feature space, a dataset-based approach to explanation has the advantage of being a coherent echo of the techniques used in ML. In the literature, the dataset may be, for example, the one

used to train the ML model or any other sample of instances which are deemed general enough to represent the whole feature space. The most studied type of explanations is *sufficient reasons*, called also *abductive explanations*. An abductive explanation for assigning a class to a given instance is a subset of the instance, viewed as a set of (feature,value) pairs, which is sufficient to guarantee the class.

The current literature has revealed that the sample-based approach is quite challenging as it faces various issues. Indeed, it has been shown in [1, 14] that dataset-based abductive explanations can be globally inconsistent in the following technical sense: two explanations for two instances corresponding to distinct classes may be compatible, meaning that a third instance exists satisfying the two sufficient reasons for contradictory classes. According to [3], it is also not possible to define an explanation function that generates subset-minimal sufficient reasons and guarantees existence of explanations (*success*) and their global consistency. Consequently, in the same paper the authors introduced various functions that guarantee consistency at the cost of success while the functions defined in [14] ensure success and violate consistency. However, there is no explanation function in the literature which generates dataset-based abductive explanations while guaranteeing both properties.

Following an axiomatic approach, we provide in this paper a **complete** investigation of functions that generate feature-based explanations from datasets. We start by proposing formal properties of explainers, then identify all the families of explainers that may be defined, shed light on the pros/cons of each family and on the links between the families. The findings unravel the origins of the above issues and uncover the full landscape of sample-based explainers. More precisely, the contributions of the paper are sixfold:

- We propose a set of axioms - desirable properties - that explainers may satisfy. Three axioms are borrowed from the literature while others are new.
- We delve into the relationships between the axioms and shed light on problematic incompatibilities of some of them. In particular, we show that the above-mentioned consistency property is incompatible with subset-minimality and success.
- We fully characterize the entire family of explainers that generate weak abductive explanations.
- We also characterize all its sub-families that satisfy subsets of compatible properties.
- We introduce the **first** (broad family of) explainers that guarantee explanations and their global consistency.
- We discuss some of its **tractable** instances including the *irrefutable explainer* and the *surrogate explainer* which is based on surrogate models.

The paper is organized as follows: Section 2 introduces sample-

\* Corresponding Author. Email: cooper@irit.fr

based explainers, a list of axioms, their links, and a characterization of the main family of explainers that satisfy two key axioms. Section 3 presents three sub-families of explainers that satisfy large subsets of compatible axioms, Section 4 discusses three tractable examples of coherent explainers, and Section 5 illustrates them on a real-world example. Section 6 discusses the obtained results, and the last section is devoted to related work and some concluding remarks. All proofs are given in the appendices.

## 2 Sample-based Explainers

Throughout the paper, we consider a *classification theory* as a tuple  $\mathbb{T} = \langle \mathbb{F}, \text{dom}, \mathbb{C} \rangle$  comprising a finite set  $\mathbb{F}$  of *features*, a function  $\text{dom}$  which returns the *domain* of every feature, where  $\text{dom}(\cdot)$  is finite and  $|\text{dom}(\cdot)| > 1$ , and a finite set  $\mathbb{C}$  of *classes* with  $|\mathbb{C}| \geq 2$ . A *literal* in  $\mathbb{T}$  is a pair  $(f, v)$  where  $f \in \mathbb{F}$  and  $v \in \text{dom}(f)$ . A set  $L$  of literals is *consistent* iff, for any two elements  $(f, v)$  and  $(f', v')$  of  $L$ , if  $f = f'$ , then  $v = v'$ . A *partial assignment* is any set of literals with each feature in  $\mathbb{F}$  occurring at most once; it is called an *instance* when every feature appears once. Notice that partial assignments and instances are consistent. We denote by  $\mathbb{E}(\mathbb{T})$  the set of all possible partial assignments and by  $\mathbb{F}(\mathbb{T})$  the *feature space*, i.e., the set of all instances, of theory  $\mathbb{T}$ . We consider a *classifier* on a theory  $\mathbb{T}$  as a function  $\kappa : \mathbb{F}(\mathbb{T}) \rightarrow \mathbb{C}$ , i.e., mapping every instance in  $\mathbb{F}[\mathbb{T}]$  to a class in the set  $\mathbb{C}$  of classes.

In this paper, we are interested in explaining decisions taken by a classifier  $\kappa$  for instances of a theory  $\mathbb{T}$ . Explanations are generated using a **subset** of the feature space. For the sake of generality, its origin is left unspecified and it may be any sample of instances.

**Definition 1.** A question is a tuple  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  such that  $\mathbb{T}$  is a classification theory,  $\kappa$  is a classifier on  $\mathbb{T}$ ,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $x \in \mathcal{D}$ .

An *explainer* is a function that takes as input a question and outputs a set of explanations. In the paper, we focus on *feature-based explanations* which describe the input features that contribute to a classifier's output for a given instance. Such explanations are thus partial assignments of the theory.

**Definition 2.** An explainer is a function  $\mathbf{L}$  mapping every question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  into a subset of  $\mathbb{E}(\mathbb{T})$ . Every  $E \in \mathbf{L}(\mathbf{Q})$  is an explanation of  $\kappa(x)$  under the dataset  $\mathcal{D}$ .

To theoretically scrutinize such explainers, we propose axioms (or formal properties) that they may satisfy. Axioms are important not only for a better understanding of the explanation process in general, but also for clarifying the basic assumptions underlying explainers, highlighting pros and cons of an explainer, comparing different (families of) explainers, and for also identifying families of explainers that have not been explored yet. We provide in Table 1 a list of ten, of which the three first ones (Success, Strong Irreducibility, Coherence) have counterparts in [14] while the others are **new**. *Success* guarantees at least one explanation to every question. *(Strong) Irreducibility* ensures that explanations do not contain unnecessary information to the explained decision. *Irreducibility* is tailored to sample-based explainers while its strong version concerns explainers that use the feature space. *Coherence* is crucial for sample-based explainers as it ensures global consistency of all generated explanations. Technically, two explanations for two instances corresponding to distinct classes should not be compatible, otherwise there would exist a third instance satisfying the two sufficient reasons for contradictory classes. Let us now introduce the new axioms. As we are

interested in feature-based explanations, *Feasibility* states that an explanation should be part of the instance being explained. *Validity* ensures that generated explanations are locally consistent within the dataset. *(Strong) Completeness* ensures that no valid explanations are omitted. *Monotonicity* states that the process of constructing explanations is monotonic, that is, an explanation remains valid if the dataset is enlarged. *Counter-monotonicity* (CM) states that enlarging a dataset can only lead to discarding explanations.

We show that some axioms follow from others. Despite these dependencies, we keep all the axioms because some explainers may satisfy a property but violate some of those from which it follows. Hence, they are useful for discriminating explainers.

**Proposition 1.** The following implications hold.

- *Completeness*  $\Rightarrow$  *Strong Completeness*,  
*Strong Completeness*  $\Rightarrow$  *Success*.
- *Irreducibility*  $\Rightarrow$  *Strong Irreducibility*.
- *Success, Feasibility and Coherence*  $\Rightarrow$  *Validity*.
- *Feasibility, Validity and Completeness*  $\Rightarrow$  *Counter-Monotonicity*.

Despite the importance of all the axioms, some of them are incompatible, i.e., they cannot be satisfied all together by an explainer.

**Theorem 1.** The axioms of every set  $\mathbf{I}_{i=1,5}$  below are incompatible.

- ( $\mathbf{I}_1$ ) *Feasibility, Success, Coherence and Irreducibility*.
- ( $\mathbf{I}_2$ ) *Feasibility, Coherence and Completeness*.
- ( $\mathbf{I}_3$ ) *Strong Irreducibility and Strong Completeness*.
- ( $\mathbf{I}_4$ ) *Feasibility, Validity, Success, Irreducibility and Monotonicity*.
- ( $\mathbf{I}_5$ ) *Feasibility, Validity, Success, Irreducibility and CM*.

Note that the important property of Irreducibility is incompatible with most of key axioms like Success and Coherence. This shows that subset-minimality of explanations has a cost in the sample-based explaining context. Coherence (or global consistency of explanations) cannot be ensured together with some other properties. Below is a list of five maximal sets of compatible axioms.

**Theorem 2.** The axioms of every set  $\mathbf{C}_{i=1,5}$  below are compatible.

- ( $\mathbf{C}_1$ ) *Feasibility, Validity, Success, (Strong) Completeness, CM*.
- ( $\mathbf{C}_2$ ) *Feasibility, Validity, Success, (Strong) Irreducibility*.
- ( $\mathbf{C}_3$ ) *Feasibility, Validity, Success, Coherence, Monotonicity, CM, Strong Completeness*.
- ( $\mathbf{C}_4$ ) *Feasibility, Validity, Success, Coherence, Monotonicity, CM, Strong Irreducibility*.
- ( $\mathbf{C}_5$ ) *Feasibility, Validity, Coherence, (Strong) Irreducibility, Monotonicity, CM*.

From the above results, we have the following characterisation.

**Theorem 3.** Considering all sets of axioms among those listed in Table 1 which include Feasibility and Validity, the only minimal incompatible sets are  $\mathbf{I}_1, \dots, \mathbf{I}_5$  and the only maximal compatible sets are  $\mathbf{C}_1, \dots, \mathbf{C}_5$ .

Recall that Feasibility defines the **(feature-based) type** of explanations and Validity guarantees their **local consistency** within the dataset. They are thus mandatory in our investigation of feature-based explanations. In what follows, we provide a full characterization of the entire family of explainers that satisfy them. We show that instances of this family generate the so-called *weak abductive explanations* [14, 30]. Such explanations are *sufficient reasons* for predicting the classes of instances of a classification theory. They are

Feasibility	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), E \subseteq x.$
Validity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), \nexists y \in \mathcal{D} \text{ s.t. } E \subseteq y \text{ and } \kappa(y) \neq \kappa(x).$
Success	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \mathbf{L}(\mathbf{Q}) \neq \emptyset.$
Coherence	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle \text{ and } \mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}, x' \rangle \text{ s.t. } \kappa(x) \neq \kappa(x'),$ then $\forall E \in \mathbf{L}(\mathbf{Q}), \forall E' \in \mathbf{L}(\mathbf{Q}'), E \cup E'$ is inconsistent.
Irreducibility	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), \forall l \in E, \exists x' \in \mathcal{D} \text{ s.t. } \kappa(x') \neq \kappa(x) \text{ and } E \setminus \{l\} \subseteq x'.$
Strong Irreducibility	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \in \mathbf{L}(\mathbf{Q}), \forall l \in E, \exists x' \in \mathbb{F}(\mathbb{T}) \text{ s.t. } \kappa(x') \neq \kappa(x) \text{ and } E \setminus \{l\} \subseteq x'.$
Completeness	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \subseteq x, \text{ if } E \notin \mathbf{L}(\mathbf{Q}), \text{ then } \exists y \in \mathcal{D} \text{ s.t. } E \subseteq y \text{ and } \kappa(y) \neq \kappa(x).$
Strong Completeness	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall E \subseteq x, \text{ if } E \notin \mathbf{L}(\mathbf{Q}), \text{ then } \exists y \in \mathbb{F}(\mathbb{T}) \text{ s.t. } E \subseteq y \text{ and } \kappa(y) \neq \kappa(x).$
Monotonicity	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall \mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle, \text{ if } \mathcal{D} \subseteq \mathcal{D}', \text{ then } \mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}(\mathbf{Q}').$
Counter-Monotonicity (CM)	$\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle, \forall \mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle, \text{ if } \mathcal{D} \subseteq \mathcal{D}', \text{ then } \mathbf{L}(\mathbf{Q}') \subseteq \mathbf{L}(\mathbf{Q}).$

Table 1. Formal Properties of Explainer  $\mathbf{L}$ .

also known as *Prime Implicants* in [7, 17, 24, 30]. In this literature, they are generated from the **whole feature space**, answering thus only the subset of questions where  $\mathcal{D} = \mathbb{F}(\mathbb{T})$ . In what follows, we present their counterpart in a sample-based setting.

**Definition 3.** Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question and  $E \in \mathbb{E}(\mathbb{T})$ .  $E$  is a weak abductive explanation (dwAXp) of  $\kappa(x)$  iff:

- $E \subseteq x$ ,
- $\forall y \in \mathcal{D}$ , if  $E \subseteq y$  then  $\kappa(y) = \kappa(x)$ .

We denote by  $\mathbf{L}_{\text{dw}}$  the explainer generating dwAXp's and by  $\mathbf{L}_{\text{w}}$  the explainer that generates dwAXp's from the feature space  $\mathbb{F}(\mathbb{T})$ .

Note that  $\mathbf{L}_{\text{w}}$  returns for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  the dwAXp of  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathbb{F}(\mathbb{T}), x \rangle$  (i.e.,  $\mathbf{L}_{\text{w}}(\mathbf{Q}) = \mathbf{L}_{\text{dw}}(\mathbf{Q}')$ ), thus neglecting  $\mathcal{D}$  and using the feature space instead.

**Example 1.** Consider the theory  $\mathbb{T}_1$  made of two binary features and a binary classifier  $\kappa_1$  that gives the predictions in the table below. Consider also the dataset  $\mathcal{D}_1 = \{x_1, x_2\}$  and let us focus on the question  $\mathbf{Q}_1 = \langle \mathbb{T}_1, \kappa_1, \mathcal{D}_1, x_1 \rangle$ . It can be checked that  $\mathbf{L}_{\text{w}}(\mathbf{Q}_1) = \{E_3\}$  while  $\mathbf{L}_{\text{dw}}(\mathbf{Q}_1) = \{E_2, E_3\}$  ( $E_2$  is not valid on  $\mathbb{F}(\mathbb{T}_1)$ ).

$\mathbb{F}(\mathbb{T}_1)$	$f_1$	$f_2$	$\kappa_1(x_i)$
$x_1$	0	0	0
$x_2$	0	1	1
$x_3$	1	0	1
$x_4$	1	1	0

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

Below we provide a characterization of the **entire family** of explainers that provide weak abductive explanations (dwAXp). We show that they are the **only ones** to satisfy Feasibility and Validity.

**Theorem 4.** An explainer  $\mathbf{L}$  satisfies Feasibility and Validity iff for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .

From a computational complexity stance, the problem of testing whether an explanation belongs to  $\mathbf{L}_{\text{w}}$  is intractable. Indeed, it has been shown in [15] that the complexity of testing whether  $E$  is a weak abductive explanation in the whole feature space is co-NP-complete. However, the problem is tractable in the sample setting (i.e., using the function  $\mathbf{L}_{\text{dw}}$ ). Indeed, it has been shown in [14] that testing the validity of sample-based weak abductive explanations is linear in the size of the sample whatever the classifier  $\kappa$ .

**Property 1.** [14] Let  $\mathbb{T} = \langle \mathbb{F}, \text{dom}, \mathbb{C} \rangle$  be a theory with  $n = |\mathbb{F}|$ ,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  with  $m = |\mathcal{D}|$  and  $E \in \mathbb{E}(\mathbb{T})$ . Testing whether  $E$  is a dwAXp can be achieved in  $O(mn)$  time.

### 3 Sub-Families of Explainers

In this section, we scrutinize the types of explainers that can be defined from the set of axioms. More precisely, we characterize all sub-families of explainers which select in different ways subsets of weak abductive explanations. Each sub-family satisfies one of the first four subsets ( $\mathbf{C}_i$ ) of compatible axioms identified in Theorem 2. The last set  $\mathbf{C}_5$  is satisfied by the trivial explainer which returns the empty set for each question. Furthermore, existing explainers from [3] satisfy Coherence and Irreducibility at the cost of Success.

#### 3.1 Weak Abductive Explainers

The first sub-family contains two explainers, namely the function  $\mathbf{L}_{\text{dw}}$  which returns **all** weak abductive explanations and  $\mathbf{L}_{\text{w}}$  which explores all the feature space. We show that  $\mathbf{L}_{\text{dw}}$  privileges **Completeness** in the two conflicts ( $\mathbf{I}_2$ ) and ( $\mathbf{I}_3$ ). Furthermore, it is the **only** explainer that satisfies Feasibility, Validity and Completeness.

**Theorem 5.** An explainer  $\mathbf{L}$  satisfies Feasibility, Validity and Completeness iff  $\mathbf{L} = \mathbf{L}_{\text{dw}}$ .

We summarize below the complete list of axioms satisfied and violated respectively by the function  $\mathbf{L}_{\text{dw}}$ . Notice that  $\mathbf{L}_{\text{dw}}$  satisfies the first largest set ( $\mathbf{C}_1$ ) of compatible axioms from Theorem 2.

**Theorem 6.** The explainer  $\mathbf{L}_{\text{dw}}$  satisfies Feasibility, Validity, Success, (Strong) Completeness and CM. It violates the remaining axioms.

The explainer  $\mathbf{L}_{\text{dw}}$  satisfies thus all the axioms that are compatible with Completeness except Monotonicity. The latter is violated due to incompleteness of information of a dataset. If the latter is enlarged, it is possible that a dwAXp will no longer be valid as the new dataset would contain a new instance which invalidates the explanation (violation of the second condition of Def. 3).

**Example 1 (Cont)** Consider the dataset  $\mathcal{D}_2 = \{x_1, x_2, x_3\}$ . Note that  $E_2 \notin \mathbf{L}_{\text{dw}}(\mathbf{Q}_2)$ , where  $\mathbf{Q}_2 = \langle \mathbb{T}_1, \kappa_1, \mathcal{D}_2, x_1 \rangle$ .

The second function  $\mathbf{L}_{\text{w}}$  which generates weak abductive explanation from the feature space is the **only** explainer that satisfies the four axioms Feasibility, Validity, Monotonicity and Strong Completeness.

**Theorem 7.** An explainer  $\mathbf{L}$  satisfies Feasibility, Validity, Monotonicity and Strong Completeness iff  $\mathbf{L} = \mathbf{L}_{\text{w}}$ .

Unlike  $\mathbf{L}_{\text{dw}}$ , the explainer  $\mathbf{L}_{\text{w}}$  satisfies Coherence and all the other axioms of the set  $\mathbf{C}_3$ .

**Theorem 8.** The explainer  $\mathbf{L}_{\text{w}}$  satisfies Feasibility, Validity, Success, Coherence, Monotonicity, CM and Strong Completeness. It violates (Strong) Irreducibility and Completeness.

The violation of Completeness by  $\mathbf{L}_w$  is due to the fact that the axiom is tailored for sample-based explainers. It is also worth noticing that since  $\mathbf{L}_{dw}$  satisfies Counter-Monotonicity, it recovers any explanation that is generated by  $\mathbf{L}_w$  from the entire feature space of the theory at hand, i.e., it finds all the “real” explanations under any dataset (see the Fidelity property in Table 2).

However, the function  $\mathbf{L}_{dw}$  may generate additional explanations which are valid only in the dataset. These new explanations are the main culprits for the violation of Coherence by  $\mathbf{L}_{dw}$ .

### 3.2 Concise Abductive Explainers

Our next characterization concerns explainers that satisfy **Irreducibility** in the four conflicts ( $\mathbf{I}_1, \mathbf{I}_3, \mathbf{I}_4, \mathbf{I}_5$ ). We show that they generate subset-minimal weak abductive explanations. Note that such explainers have been studied in [1, 14].

**Definition 4.** Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question and  $E \in \mathbb{E}(\mathbb{T})$ .  $E$  is a concise abductive explanation (cAXp) of  $\kappa(x)$  iff:

- $E \in \mathbf{L}_{dw}(\mathbf{Q})$ ,
- $\nexists E' \in \mathbf{L}_{dw}(\mathbf{Q})$  such that  $E' \subset E$ .

We denote by  $\mathbf{L}_{dc}$  the explainer generating cAXp.

**Example 1 (Cont)**  $\mathbf{L}_{dc}(\mathbf{Q}_1) = \{E_2\}$  and  $\mathbf{L}_{dw}(\mathbf{Q}_1) = \{E_2, E_3\}$ .

We show next that explainers that provide cAXp’s are the **only ones** satisfying Feasibility, Validity, and Irreducibility.

**Theorem 9.** An explainer  $\mathbf{L}$  satisfies Feasibility, Validity, and Irreducibility iff for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{dc}(\mathbf{Q})$ .

An instance of this family of explainers is  $\mathbf{L}_{dc}$  which generates **all** the subset-minimal weak abductive explanations. We show that it satisfies the second set ( $\mathbf{C}_2$ ) of compatible axioms, and thus violates several properties due to their incompatibility with Irreducibility.

**Theorem 10.** The explainer  $\mathbf{L}_{dc}$  satisfies Feasibility, Validity, Success and (Strong) Irreducibility. It violates the remaining axioms.

**Property 2.** Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question.

- $x \in \mathbf{L}_{dw}(\mathbf{Q})$ .
- $\mathbf{L}_{dc}(\mathbf{Q}) \subseteq \mathbf{L}_{dw}(\mathbf{Q})$ .

Another explainer which generates concise abductive explanations is  $\mathbf{L}_c$  which explores the feature space. Such explanations, known as prime implicants, have been largely studied in the literature (e.g., [5, 6, 8, 15, 17, 18, 24, 25, 36]).

**Definition 5.** A prime implicant explainer is a function  $\mathbf{L}_c$  such that for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}_c(\mathbf{Q}) = \mathbf{L}_{dc}(\mathbf{Q}')$ , where  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathbb{F}(\mathbb{T}), x \rangle$ .

The function  $\mathbf{L}_c$  satisfies all the axioms of the set ( $\mathbf{C}_4$ ), hence all axioms except (Strong) Completeness, and also Irreducibility despite the fact it generates subset-minimal explanations. This violation is merely due to the definitions of the axiom which is sample-based.

**Theorem 11.** The function  $\mathbf{L}_c$  satisfies all the axioms except Irreducibility and (Strong) Completeness.

**Remark:** Note that the inclusion  $\mathbf{L}_c(\mathbf{Q}) \subseteq \mathbf{L}_{dc}(\mathbf{Q})$  may not hold.

**Example 1 (Cont)**  $\mathbf{L}_{dc}(\mathbf{Q}_1) = \{E_2\}$  while  $\mathbf{L}_c(\mathbf{Q}_1) = \{E_3\}$ .

From a computational complexity perspective, the problems of *testing* and *finding* one explanation of  $\mathbf{L}_c$  are intractable. Indeed, it has been shown in [16] that the complexity of testing whether a partial assignment  $E$  is an explanation is in  $\mathbf{P}^{\text{NP}}$  (and is NP-hard), and the complexity of finding one explanation is in  $\mathbf{FP}^{\text{NP}}$ . However, the two problems are tractable when  $\mathbf{L}_{dc}$  is used [14].

**Property 3.** [14] Let  $\mathbb{T} = \langle \mathbb{F}, \text{dom}, \mathbb{C} \rangle$  be a theory with  $n = |\mathbb{F}|$ ,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  with  $m = |\mathcal{D}|$  and  $E \in \mathbb{E}(\mathbb{T})$ . Finding a cAXp can be achieved in  $O(mn^2)$  time.

### 3.3 Coherent Sample-based Abductive Explainers

The two previous sub-families of explainers, in particular their sample-based instances ( $\mathbf{L}_{dw}, \mathbf{L}_{dc}$ ) privilege Completeness and Irreducibility at the cost of Coherence, hence their explanations may be globally inconsistent. Indeed, we have seen that Coherence is incompatible with completeness, and with the pair (Success, Irreducibility). In what follows, we propose the **first (family of) explainers** that satisfy Success and Coherence. We start by showing that its instances generate weak abductive explanations, meaning that they select a collection of globally consistent dwAXp’s.

**Theorem 12.** If an explainer  $\mathbf{L}$  satisfies Feasibility, Success and Coherence, then for any question  $\mathbf{Q}$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{dw}(\mathbf{Q})$ .

Before introducing the novel family, let us first provide some useful notions, including a *coherent set* of partial assignments. The latter is nothing more than a set which satisfies the consistency condition of the Coherence axiom.

**Definition 6.** Let  $\mathbb{T}$  be a theory,  $X \subseteq \mathbb{E}(\mathbb{T})$ ,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  and  $\kappa$  a classifier.  $X$  is coherent under  $(\mathcal{D}, \kappa)$  iff  $\nexists E, E' \in X$  such that:

- $E \cup E'$  is consistent, and
- $\exists y, z \in \mathcal{D}$  such that  $E \subseteq y$ ,  $E' \subseteq z$ , and  $\kappa(y) \neq \kappa(z)$ .

Otherwise,  $X$  is said to be incoherent.

**Example 1 (Cont)** Let  $E_1 = \{(f_1, 0)\}$  and  $E_2 = \{(f_2, 0)\}$ . The set  $\{E_1, E_2\}$  is incoherent since  $E_1 \cup E_2$  is consistent,  $E_1 \subseteq x_2$ ,  $E_2 \subseteq x_1$  and  $\kappa_1(x_2) \neq \kappa_1(x_1)$ .

Since no two distinct instances of a feature space are consistent with each other, every dataset is coherent.

**Property 4.** For any  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  and any classifier  $\kappa$ ,  $\mathcal{D}$  is coherent under  $(\mathcal{D}, \kappa)$ .

Let us introduce the notion of *envelope*, which is a coherent set of weak abductive explanations covering every instance of a dataset.

**Definition 7.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ ,  $\kappa$  a classifier. An envelope under  $(\mathcal{D}, \kappa)$  is any  $X \subseteq \mathbb{E}(\mathbb{T})$  such that the following hold:

- $X$  is coherent under  $(\mathcal{D}, \kappa)$ ,
- $\forall E \in X, \exists x \in \mathcal{D}$  such that  $E \in \mathbf{L}_{dw}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ ,
- $\forall x \in \mathcal{D}, \exists E \in X$  such that  $E \subseteq x$ .

Let  $\text{Coh}(\mathcal{D}, \kappa)$  be the set of all envelopes under  $(\mathcal{D}, \kappa)$ .

It is easy to see that the set  $\text{Coh}(\mathcal{D}, \kappa)$  is not empty as it contains at least the dataset  $\mathcal{D}$  itself. Furthermore, each  $X \in \text{Coh}(\mathcal{D}, \kappa)$  contains a subset of the dwAXp’s of every instance in  $\mathcal{D}$ .

**Proposition 2.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier.

- Let  $X \in \text{Coh}(\mathcal{D}, \kappa)$ . For any  $x \in \mathcal{D}$ , for any  $E \in X$ , if  $E \subseteq x$ , then  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ .
- $\mathcal{D} \in \text{Coh}(\mathcal{D}, \kappa)$ .
- For any  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ ,  $\exists X \in \text{Coh}(\mathcal{D}, \kappa)$  such that  $X$  is a subset-maximal envelope of  $\text{Coh}(\mathcal{D}, \kappa)$  and  $E \in X$ .

**Example 1 (Cont)**  $\text{Coh}(\mathcal{D}_1, \kappa_1)$  contains the following envelopes:

- $X_1 = \mathcal{D}_1 = \{x_1, x_2\}$ ,
- $X_2 = \{\{(f_2, 0)\}, \{(f_2, 1)\}\}$ ,
- $X_3 = \{\{(f_2, 0)\}, x_2\}$ ,
- $X_4 = \{\{(f_2, 1)\}, x_1\}$ ,
- $X_5 = \{x_1, x_2, \{(f_2, 0)\}, \{(f_2, 1)\}\}$ .

Note that, for example, the set  $\{\{(f_2, 0)\}\}$  is not an envelope since  $\{(f_2, 0)\}$  is not a  $\text{dwAXp}$  of the instance  $x_2$ .

Let us now define the novel family of *coherent explainers* whose instances return, under any dataset, an envelope.

**Definition 8.** A coherent explainer is a function  $\mathbf{L}_{\text{co}}$  such that for any  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ , the following hold:

- $\mathbf{L}_{\text{co}}(\mathbf{Q}) = \{E \in X \mid E \subseteq x\}$ , where  $X = \bigcup_{z \in \mathcal{D}} \mathbf{L}_{\text{co}}(\langle \mathbb{T}, \kappa, \mathcal{D}, z \rangle)$ ,
- $X \in \text{Coh}(\mathcal{D}, \kappa)$ .

We show next that coherent explainers are the **only ones** to satisfy together the three axioms Success, Feasibility and Coherence, hence to guarantee an explanation to every instance as well as global consistency of explanations.

**Theorem 13.** An explainer  $\mathbf{L}$  satisfies Feasibility, Success and Coherence *iff*  $\mathbf{L}$  is a coherent explainer.

In addition to the three properties (Success, Feasibility, Coherence), any coherent explainer satisfies Validity but violates Irreducibility and Completeness.

**Theorem 14.** Any coherent explainer satisfies Feasibility, Validity, Success, Coherence. It violates Irreducibility and Completeness.

The sub-family of coherent explainers is very **broad** and encompasses various explainers including the two functions  $\mathbf{L}_{\text{w}}$  and  $\mathbf{L}_{\text{c}}$ . Some of its instances are **tractable** as we will see in the next section.

## 4 Tractable Coherent Explainers

The family of coherent explainers is large and covers a variety of functions because from a given dataset  $\mathcal{D}$ , one may generate several envelopes, i.e.,  $|\text{Coh}(\mathcal{D}, \kappa)| \geq 1$ . In this section, we discuss three functions whose explanations can be generated in polynomial time.

### 4.1 Trivial Explanation Function

The first function, called *trivial explainer*, is the one whose envelope is the dataset itself. We have seen in Proposition 2 that any dataset is an envelope. This function assigns thus a single explanation to every question, which is nothing more than the instance being explained.

**Definition 9.** A trivial explainer is a function  $\mathbf{L}_{\text{tr}}$  such that for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}_{\text{tr}}(\mathbf{Q}) = \{x\}$ .

The function  $\mathbf{L}_{\text{tr}}$  is clearly a coherent explainer.

**Property 5.** The function  $\mathbf{L}_{\text{tr}}$  is a coherent explainer.

**Theorem 15.** The function  $\mathbf{L}_{\text{tr}}$  satisfies all the axioms except (Strong) Irreducibility and (Strong) Completeness.

Obviously, an explainer which returns an instance as explanation of its outcome is not informative and definitely not useful for users.

### 4.2 Irrefutable Explanation Functions

We now introduce the novel *irrefutable explainer*. It uses the so-called *irrefutable envelope* that contains all non-conflicting weak abductive explanations.

**Definition 10.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ ,  $\kappa$  a classifier, and  $\mathbb{W} = \bigcup_{x \in \mathcal{D}} \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ . An irrefutable envelope under  $(\mathcal{D}, \kappa)$  is  $X \subseteq \mathbb{E}(\mathbb{T})$  such that the following hold:

- $X \subseteq \mathbb{W}$ ,
- $\forall E \in X, \nexists E' \in \mathbb{W}$  such that  $\{E, E'\}$  is incoherent,
- $\nexists X' \supset X$  that satisfies the above conditions.

The next result shows that the irrefutable envelope is unique.

**Property 6.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier. For all  $X, X' \subseteq \mathbb{E}(\mathbb{T})$ , if  $X$  and  $X'$  are irrefutable envelopes under  $(\mathcal{D}, \kappa)$ , then  $X = X'$ .

**Notation:** Throughout the paper,  $\text{Irr}(\mathcal{D}, \kappa)$  denotes the irrefutable envelope under  $(\mathcal{D}, \kappa)$  in theory  $\mathbb{T}$ .

An irrefutable envelope contains the whole dataset under which it is defined. Furthermore, it is the intersection of all subset-maximal envelopes.

**Proposition 3.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier.

- $\mathcal{D} \subseteq \text{Irr}(\mathcal{D}, \kappa)$ .
- $\text{Irr}(\mathcal{D}, \kappa) \in \text{Coh}(\mathcal{D}, \kappa)$ .
- $\text{Irr}(\mathcal{D}, \kappa) = \bigcap_{X_i \in \mathcal{S}} X_i$ , where  $\mathcal{S}$  is the set of all subset-maximal envelopes of  $\text{Coh}(\mathcal{D}, \kappa)$ .

We are now ready to present the instance of the family of coherent explainers that is based on irrefutable envelopes.

**Definition 11.** An irrefutable explainer is a function  $\mathbf{L}_{\text{ir}}$  such that for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,

$$\mathbf{L}_{\text{ir}}(\mathbf{Q}) = \{E \in \text{Irr}(\mathcal{D}, \kappa) \mid E \subseteq x\}.$$

**Example 1 (Cont)** Recall that  $\text{Coh}(\mathcal{D}_1, \kappa_1) = \{X_i \mid i = 1, \dots, 5\}$ .  $\text{Irr}(\mathcal{D}_1, \kappa_1) = X_5$ , hence  $\mathbf{L}_{\text{ir}}(\mathbf{Q}_1) = \{x_1, \{(f_2, 0)\}\}$ .

**Example 2.** Consider the theory  $\mathbb{T}_2$  made of two binary features and a binary classifier  $\kappa_2$  which provides the predictions below for the three instances in  $\mathcal{D}_2$ . Let  $\mathbf{Q}_i = \langle \mathbb{T}_2, \kappa_2, \mathcal{D}_2, x_i \rangle$ , with  $i \in \{1, 2, 3\}$ .

$\mathcal{D}_2$	$f_1$	$f_2$	$\kappa_2(x_i)$	
$x_1$	0	0	0	$E_1 = \{(f_1, 0)\}$
$x_2$	1	0	0	$E_2 = \{(f_2, 0)\}$
$x_3$	1	1	1	$E_3 = \{(f_2, 1)\}$
				$E_4 = x_1 \quad E_5 = x_2 \quad E_6 = x_3$

Note that  $\bigcup_{i=1}^3 \mathbf{L}_{\text{dw}}(\mathbf{Q}_i) = \{E_1, \dots, E_6\}$ . Note also that the set  $\{E_1, E_3\}$  is incoherent and  $\text{Irr}(\mathcal{D}_2, \kappa_2) = \{E_2, E_4, E_5, E_6\}$ . So  $\mathbf{L}_{\text{ir}}(\mathbf{Q}_1) = \{E_2, E_4\}$ ,  $\mathbf{L}_{\text{ir}}(\mathbf{Q}_2) = \{E_2, E_5\}$ ,  $\mathbf{L}_{\text{ir}}(\mathbf{Q}_3) = \{E_6\}$ .

**Property 7.**  $\mathbf{L}_{\text{ir}}$  is a coherent explainer, and for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $x \in \mathbf{L}_{\text{ir}}(\mathbf{Q})$ .

Unlike the trivial explainer,  $\mathbf{L}_{\text{ir}}$  violates Monotonicity and Counter-Monotonicity.

**Theorem 16.** *The function  $\mathbf{L}_{\text{ir}}$  satisfies Feasibility, Validity, Success and Coherence. It violates all the other axioms.*

**Remark:** Unlike  $\mathbf{L}_{\text{dw}}$ , the coherent explainers  $\mathbf{L}_{\text{tr}}$  and  $\mathbf{L}_{\text{ir}}$  do not necessarily recover the explanations that are generated from the feature space, i.e., the following inclusion may not hold:  $\mathbf{L}(\mathbf{Q}') \subseteq \mathbf{L}(\mathbf{Q})$ , for  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathbb{F}(\mathbb{T}), x \rangle$ . See the Fidelity property in Table 2.

In addition to guaranteeing the existence of explanations and their global consistency, the irrefutable explainer is tractable. Indeed, a subset-minimal explanation can be generated in polynomial time by a greedy algorithm [13] based on the following theorem.

**Theorem 17.** *Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question, where  $\kappa$  can be evaluated in polynomial time. Testing whether  $E \in \mathbf{L}_{\text{ir}}(\mathbf{Q})$  can be achieved in polynomial time.*

### 4.3 Surrogate Classifier

In this section we describe another type of coherent explainers. Starting from a result in [2] stating that Coherence is ensured by explainers that provide sufficient reasons from the whole feature space, the idea is, given a question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ , to find a *surrogate classifier*  $\sigma$  which is equal to  $\kappa$  on  $\mathcal{D}$  but which allows tractable explaining on the whole feature space. Explanations of  $\mathbf{Q}$  are then explanations of  $\mathbf{Q}' = \langle \mathbb{T}, \sigma, \mathbb{F}(\mathbb{T}), x \rangle$ . Tractability follows if we choose  $\sigma$  from a family of classifiers that allows tractable explaining on the whole feature space [4, 12, 22, 27].

The approach relies heavily on  $\sigma$ , thus the question of its **existence** naturally arises. The answer is fortunately positive. Indeed, it is always possible to find a *decision tree*  $\sigma$  for any model  $\kappa$  with 100% accuracy on any dataset  $\mathcal{D}$  (by over-fitting, if necessary). Each instance  $x \in \mathcal{D}$  corresponds to a unique path from the root to a leaf of the decision tree and will be consulted, during construction of the decision tree, at each node of the path (and at no other node). In the case when all features are boolean, the length of such a path is at most  $n$ , the number of features, under the reasonable assumption that the same boolean feature is not redundantly tested two times on the same path. Standard heuristic algorithms [10, 39] which choose which feature to branch on at each node according to a score, such as entropy, will thus have a complexity of  $O(n^2m)$ , where  $m$  is the number of instances in  $\mathcal{D}$ .

**Proposition 4.** *Let  $\mathbb{T}$  be a theory and  $\mathcal{D} \subseteq \mathbb{T}$ . For any classifier  $\kappa$ , there exists a decision-tree classifier  $\sigma$  on  $\mathbb{T}$  such that  $\forall y \in \mathcal{D}$ ,  $\sigma(y) = \kappa(y)$ .*

We now introduce the novel coherent explainers, called *surrogate explainers*, which generate weak abductive explanations from the entire feature space but using the predictions of a surrogate classifier.

**Definition 12.** *Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{T}$ ,  $\kappa$  and  $\sigma$  two classifiers on  $\mathbb{T}$  such that  $\forall y \in \mathcal{D}$ ,  $\sigma(y) = \kappa(y)$ . A surrogate explainer of  $\kappa$  is a function  $\mathbf{L}_{\text{su}}$  such that  $\forall x \in \mathcal{D}$ ,*

$$\mathbf{L}_{\text{su}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle) = \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \sigma, \mathbb{F}(\mathbb{T}), x \rangle).$$

$\mathbf{L}_{\text{DT}}$  denotes the surrogate explainer where  $\sigma$  is a decision tree.

**Remark:** Notice that the same surrogate classifier  $\sigma$  is used to explain the predictions of  $\kappa$  for all instances of a given dataset. However, since  $\sigma$  and  $\kappa$  may disagree on instances outside a dataset,  $\mathbf{L}_{\text{su}}$  may use different surrogate classifiers for different datasets.

We show next that it is **tractable** to find a surrogate explainer  $\mathbf{L}_{\text{DT}}$  which is grounded on a decision tree. The decision tree can be converted into a set of mutually-exclusive decision rules [32] whose premises can be viewed as an explanation for all instances associated with the corresponding leaf. The path from the root to a leaf is, however, often redundant as an explanation [27]. However, in polynomial time it is possible to construct an explainer that is complete on  $\mathbb{F}$ , satisfies coherence and such that explanations are weak abductive explanations of the decision tree.

**Theorem 18.** *A surrogate explainer  $\mathbf{L}_{\text{DT}}$  can be found in polynomial time.*

Recall that  $\kappa$  is a black-box function. Thus finding explanations valid over the whole feature space  $\mathbb{F}(\mathbb{T})$  requires exponential time. Furthermore, the set of dwAXp's of  $\kappa$  over  $\mathcal{D}$  do not satisfy Coherence. As we have just seen, one way to achieve tractability while guaranteeing Coherence is to consider the explanations over  $\mathbb{F}(\mathbb{T})$  of a surrogate function  $\sigma$ . The fact that  $\sigma$  and  $\kappa$  agree on  $\mathcal{D}$  means that they have the same dwAXp's. Of course, since we make what is essentially an arbitrary choice of the function  $\sigma$ , rather than studying  $\kappa$  over the whole feature space, it is clear that the explanations returned do not necessarily correspond to explanations of  $\kappa$  over  $\mathbb{F}(\mathbb{T})$ . On the other hand, we show that for *any* envelope  $X$ , there exists a surrogate function  $\sigma_X$  (expressible as a decision list) such that for all  $x \in \mathcal{D}$ , any  $E \in X$  such that  $E \subseteq x$  is a dwAXp of  $\langle \mathbb{T}, \sigma_X, \mathbb{F}(\mathbb{T}), x \rangle$ .

**Proposition 5.** *Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ ,  $\kappa$  a classifier and  $X$  an envelope under  $(\mathcal{D}, \kappa)$ . There exists a surrogate function  $\sigma_X : \mathbb{F}(\mathbb{T}) \rightarrow \mathbb{C}$  such that (1)  $\sigma_X$  and  $\kappa$  are equal on  $\mathcal{D}$  and (2) for all  $x \in \mathcal{D}$ ,  $E \in X$  and  $E \subseteq x$  implies  $E$  is a dwAXp of  $\langle \mathbb{T}, \sigma_X, \mathbb{F}(\mathbb{T}), x \rangle$ .*

Observe that  $\sigma_X$  (defined in the proof of Proposition 5) can be described by a decision list containing at most  $m$  rules (one for each instance in  $\mathcal{D}$ ). Of course, in practice, there may be many redundant rules that can be eliminated from this list. We could also express  $\sigma_X$  as a decision tree or as a d-DNNF, a language of boolean functions which also allows tractable explaining [22].

Let us now give the list of axioms satisfied by surrogate explainers and hence show that they are coherent explainers.

**Theorem 19.** *The explainer  $\mathbf{L}_{\text{su}}$  satisfies Feasibility, Validity, Success and Coherence. It violates the remaining axioms.*

## 5 Worked Example

As an example, we study the well-known zoo dataset [38] which consists of 101 instances corresponding to animals from a zoo. Each animal is described by 16 features and belongs to one of 7 classes: Mammal, Bird, Reptile, Fish, Amphibian, Bug or Invertebrate. For example, *antelope* is a mammal and *crow* is a bird. We assume that the classifier  $\kappa$  correctly classifies all 101 instances in the dataset.

In the case of a black-box classifier  $\kappa$  over a large feature space, we cannot perform an exhaustive search over all instances. First consider concise abductive explanations based on this dataset (cAXp's). A cAXp of *antelope* is  $\{(\text{milk}, 1)\}$ , since all other animals in the dataset that give milk are also mammals, and a cAXp of *crow* is  $\{(\text{feathers}, 1)\}$ , since all other animals in the dataset with feathers are also birds. Although no animal in the dataset milks its young and has feathers, these explanations are consistent and hence  $\mathbf{L}_{\text{dc}}$  (same for  $\mathbf{L}_{\text{dw}}$ ) does not satisfy coherence.

If we now look for coherent explanations, we find a size-14 irrefutable explanation for classifying *antelope* as a mammal, namely

Axioms / Explainers	$L_w$	$L_c$	$L_{dw}$	$L_{dc}$	$L_{co}$	$L_{tr}$	$L_{ir}$	$L_{su}$
Feasibility	✓	✓	✓	✓	✓	✓	✓	✓
Validity	✓	✓	✓	✓	✓	✓	✓	✓
Success	✓	✓	✓	✓	✓	✓	✓	✓
Coherence	✓	✓	×	×	✓	✓	✓	✓
Irreducibility	×	×	×	✓	×	×	×	×
Strong Irreducibility	×	✓	×	✓	—	×	×	×
Completeness	×	×	✓	×	×	×	×	×
Strong Completeness	✓	×	✓	×	—	×	×	×
Monotonicity	✓	✓	×	×	—	✓	×	×
Counter-Monotonicity	✓	✓	✓	×	—	✓	×	×
Fidelity: $L_w(Q) \subseteq L(Q)$	✓	×	✓	×	—	×	×	×

**Table 2.** The symbols ✓, × and — stand for the axiom is satisfied, violated and unknown respectively by the explainer.

{(hair,1), (feathers,0), (eggs,0), (milk,1), (airborne,0), (aquatic,0), (predator,0), (toothed,1), (backbone,1), (breathes,1), (venomous,0), (legs,4), (tail,1), (catsize,1)}. This is minimal in that all size-13 explanations obtained by deleting a literal from this explanation are *not* irrefutable. This demonstrates that there is a risk that irrefutable explanations may be very large (as in this case involving 14 of the 16 features) and hence not very informative for the user.

The solution embodied by the notion of irrefutable explanations guarantees coherence but at the expense of the size of the explanation. Another solution is the use of a surrogate classifier as discussed in Section 4.3. We applied the classic ID3 algorithm [31] to the zoo dataset to obtain a decision tree with 100% accuracy on the dataset. Since the most discriminating feature is milk which splits all mammals from all non-mammals in the dataset, this is the feature associated with the root of the decision tree found by ID3. The resulting unique AXp of *antelope* for this decision tree is simply {(milk,1)}. This size-1 explanation based on a surrogate classifier is in stark contrast with the size-14 irrefutable explanation. The explanation for *crow* is {(milk,0), (feathers,1)}, which is, of course, coherent with the explanation {(milk,1)} for *antelope*.

## 6 Discussion

This paper presented a comprehensive study of explainers that generate sufficient reasons from samples, thereby advancing our understanding of their families, strengths and weaknesses. It proposed some basic axioms, or formal properties, and showed that some subsets uniquely define various families of explainers that all generate weak abductive explanations. In other words, it identified the unique (family of) explainers that satisfy a given subset of axioms.

The reason of the diversity of families is that certain axioms are incompatible, making it tricky to define efficient sample-based functions that satisfy all the axioms. Indeed, it is not possible to define a function that returns concise (irreducible) explanations while guaranteeing at least one explanation for every instance and global coherence of explanations of all instances. Thus, one of the three properties should be abandoned. The right-to-an-explanation imposed by legislators [20, 37] impels to keep Success and the need of non-erroneous explanations suggests keeping Coherence. Thus, the compromise would be to sacrifice Irreducibility and accept explanations that contain unnecessary information. For the same reasons, it is reasonable to sacrifice Completeness in the incompatible set  $I_2$ . These choices have been made by our novel family of Coherent explainers. On the positive side, some coherent explainers ( $L_{ir}$  and  $L_{su}$ ) allow tractable explaining. The worked example showed that irrefutable explanations by  $L_{ir}$  may contain a large number of features compared to those provided by  $L_{su}$ . However,  $L_{ir}$  coincides with  $L_w$  when it uses the whole feature space, while this is not the case of

$L_{su}$  which, even if it is based on a dataset, generates weak abductive explanations from the whole feature space. Indeed, the equality  $L_{su}(Q) = L_w(Q)$  may not hold since the surrogate model  $\sigma$  may differ from the original classifier  $\kappa$  outside the dataset. Finally, as shown in Table 2, both functions violate the Fidelity property which states that an explainer would recover **all** the "real" weak-abductive explanations, i.e., those generated from the feature space. Note that the two functions recover some but not necessarily all real explanations. This is not surprising since a sample-based approach reasons under incomplete information.

## 7 Related Work and Conclusion

There are several works in the AI literature on explaining classification models. Some of them explain the inherent reasoning of models (e.g., [6, 21, 23, 36]) while others consider classifiers as black-boxes and look for possible correlations between instances and the classes assigned to them. In this second category, explanations are generally feature-based, like *sufficient reasons* (eg. [8, 17, 19, 24, 33, 34]), *contrastive/counterfactual* explanations (eg. [19, 29, 41]), or *semi-factuals* (eg. [11, 28]). Our paper fits within the second category with a focus on sufficient reasons. However, unlike most works in the literature which explore the whole feature space to generate explanations, we follow a sample-based approach limiting search on a subset of feature space. The closest works to our are those from [1, 14, 33, 34]. The authors in [33, 34] proposed the well-known LIME and Anchors models which generate explanations using a sample of instances that are in the neighborhood of the instance being explained. It has been shown in [35, 40] that they may return incorrect explanations. The origin of this deficiency has been uncovered in [1] where the author proved that it is not possible to define an explainer that provides subset-minimal abductive explanations and guarantees Success and Coherence. As a consequence of this negative result, various sample-based explainers have been proposed, some of which satisfy Coherence at the cost of Success [3] while others promote Success [14].

Our paper generalized the above negative result, introduced novel axioms and identified several other incompatible properties. It provided full characterizations of all families of explainers that can be defined, including one ( $L_{co}$ ) that ensures Success and Coherence.  $L_{co}$  is the first family of sample-based explainers that that satisfy both axioms. We proved that some of its instances are tractable.

This papers lends itself to several developments. A challenging open problem is the definition of sample-based coherent explainers that recover as many "real" explanations as possible while being tractable. A possible solution would be the use of constraints on input data in order to solve conflicts between dwAXp's. Another perspective is an axiomatic study of sample-based contrastive explanations.

## A Proofs of Properties

**Property 1.** [14] Let  $\mathbb{T} = \langle \mathbb{F}, \text{dom}, \mathbb{C} \rangle$  be a theory with  $n = |\mathbb{F}|$ ,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  with  $m = |\mathcal{D}|$  and  $E \in \mathbb{E}(\mathbb{T})$ . Testing whether  $E$  is a  $\text{dwAXp}$  can be achieved in  $O(mn)$  time.

*Proof.* Proved in [14].  $\square$

**Property 2.** [14] Let  $\mathbb{T} = \langle \mathbb{F}, \text{dom}, \mathbb{C} \rangle$  be a theory with  $n = |\mathbb{F}|$ ,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  with  $m = |\mathcal{D}|$  and  $E \in \mathbb{E}(\mathbb{T})$ . Finding a  $\text{cAXp}$  can be achieved in  $O(mn^2)$  time.

*Proof.* Proved in [14].  $\square$

**Property 3.** For any  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  and any classifier  $\kappa$ ,  $\mathcal{D}$  is coherent under  $(\mathcal{D}, \kappa)$ .

*Proof.* Follows from the fact that instances are pairwise inconsistent by definition.  $\square$

**Property 4.** The function  $\mathbf{L}_{\text{tr}}$  is a coherent explainer.

*Proof.* Straightforward from the definition.  $\square$

**Property 5.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier. For all  $X, X' \subseteq \mathbb{E}(\mathbb{T})$ , if  $X$  and  $X'$  are irrefutable envelopes under  $(\mathcal{D}, \kappa)$ , then  $X = X'$ .

*Proof.* Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier.

► Assume that  $X$  and  $X'$  are irrefutable envelopes under  $(\mathcal{D}, \kappa)$ . Suppose that  $\exists E \in X \setminus X'$ . From the definition of irrefutable envelope,  $\nexists E' \in \mathbb{W}$  such that  $\{E, E'\}$  is incoherent. Hence,  $X' \cup \{E\}$  satisfies the second condition of the same definition. This contradicts the maximality of  $X'$ . So,  $X \subseteq X'$ . In the same way, we can show that  $X' \subseteq X$  and hence  $X = X'$ .  $\square$

**Property 6.**  $\mathbf{L}_{\text{ir}}$  is a coherent explainer, and for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $x \in \mathbf{L}_{\text{ir}}(\mathbf{Q})$ .

*Proof.*  $\mathbf{L}_{\text{ir}}$  is a coherent explainer since from Proposition 3,  $\text{Irr}(\mathcal{D}, \kappa) \in \text{Coh}(\mathcal{D}, \kappa)$  and for any question  $\mathbf{Q}$  where  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}_{\text{ir}}(\mathbf{Q}) = \{E \in \text{Irr}(\mathcal{D}, \kappa) \mid E \subseteq x\}$ .

The second property follows from the first item of Proposition 3, namely the fact that  $\mathcal{D} \subseteq \text{Irr}(\mathcal{D}, \kappa)$ .  $\square$

**Property 7.** Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question.

- $x \in \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .
- $\mathbf{L}_{\text{dc}}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .

*Proof.* The properties follow straightforwardly from the definitions.  $\square$

**Property 8.** Let  $\mathbf{L}$  be an explainer that satisfies Completeness. For any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $x \in \mathbf{L}(\mathbf{Q})$ .

*Proof.* Let  $\mathbf{L}$  be an explainer that satisfies Completeness. Consider the question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $x \in \mathbf{L}(\mathbf{Q})$  and assume that  $x \notin \mathbf{L}(\mathbf{Q})$ . Completeness of  $\mathbf{L}$  implies that  $\exists y \in \mathcal{D}$  such that  $x \subseteq y$  and  $\kappa(x) \neq \kappa(y)$ . From definition of instances,  $x = y$  which is a contradiction.  $\square$

## B Proofs of Propositions

**Proposition 1.** The following implications hold.

- Completeness  $\Rightarrow$  Strong Completeness, Strong Completeness  $\Rightarrow$  Success.
- Irreducibility  $\Rightarrow$  Strong Irreducibility.
- Success, Feasibility and Coherence  $\Rightarrow$  Validity.
- Feasibility, Validity and Completeness  $\Rightarrow$  Counter-Monotonicity.

*Proof.* Let  $\mathbf{L}$  be an explainer.

► Assume that  $\mathbf{L}$  satisfies Completeness. Let  $E \subseteq x$  be such that  $E \notin \mathbf{L}(\mathbf{Q})$ . From Completeness,  $\exists y \in \mathcal{D}$  s.t.  $E \subseteq y$  and  $\kappa(y) \neq \kappa(x)$ . Thus,  $\exists y \in \mathbb{F}(\mathbb{T})$  s.t.  $E \subseteq y$  and  $\kappa(y) \neq \kappa(x)$ . So,  $\mathbf{L}$  satisfies Strong Completeness.

Assume that  $\mathbf{L}$  satisfies Strong Completeness and that  $x \notin \mathbf{L}(\mathbf{Q})$  such that  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ . From Strong Completeness,  $\exists y \in \mathbb{F}(\mathbb{T})$  such that  $x \subseteq y$  and  $\kappa(x) \neq \kappa(y)$ . By definition of instances,  $x = y$ . Since  $\kappa$  assigns a single value to every instance, it follows that  $\kappa(x) = \kappa(y)$ . Thus  $x \in \mathbf{L}(\mathbf{Q})$  and  $\mathbf{L}(\mathbf{Q}) \neq \emptyset$ .

► Strong Irreducibility follows straightforwardly from Irreducibility since  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ .

► Assume that  $\mathbf{L}$  satisfies Success, Feasibility, Coherence and violates Validity. Then, there exists a question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  such that  $\exists E \in \mathbf{L}(\mathbf{Q})$  and  $\exists y \in \mathcal{D}$  such that  $E \subseteq y$  and  $\kappa(x) \neq \kappa(y)$ . Consider now the question  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}, y \rangle$ . From Success,  $\exists E' \in \mathbf{L}(\mathbf{Q}')$ . From Feasibility,  $E' \subseteq y$  so  $E \cup E' \subseteq y$ . Since  $y$  is consistent,  $E \cup E'$  is consistent (a). From Coherence of  $\mathbf{L}$ , since  $\kappa(x) \neq \kappa(y)$ , it follows that  $E \cup E'$  is inconsistent, which contradicts (a).

► Assume that  $\mathbf{L}$  satisfies Feasibility, Validity and Completeness and violates Counter-Monotonicity. Hence, there exist two samples  $\mathcal{D}, \mathcal{D}'$  such that  $\mathcal{D} \subseteq \mathcal{D}'$ ,  $\exists x \in \mathcal{D}$  such that  $\exists E \in \mathbf{L}(\mathbf{Q}')$  while  $E \notin \mathbf{L}(\mathbf{Q})$ , with  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle$ . From Feasibility,  $E \subseteq x$ . Since  $E \notin \mathbf{L}(\mathbf{Q})$ , then from Completeness of  $\mathbf{L}$ ,  $\exists y \in \mathcal{D}$  such that  $E \subseteq y$  and  $\kappa(x) \neq \kappa(y)$  (a). Validity of  $\mathbf{L}$  ensures that  $\nexists z \in \mathcal{D}'$  such that  $E \subseteq z$  and  $\kappa(x) \neq \kappa(z)$ . Since  $\mathcal{D} \subseteq \mathcal{D}'$ , we deduce that  $y \in \mathcal{D}'$  which contradicts (a).  $\square$

**Proposition 2.** Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier.

- Let  $X \in \text{Coh}(\mathcal{D}, \kappa)$ . For any  $x \in \mathcal{D}$ , for any  $E \in X$ , if  $E \subseteq x$ , then  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ .
- $\mathcal{D} \in \text{Coh}(\mathcal{D}, \kappa)$ .
- For any  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ ,  $\exists X \in \text{Coh}(\mathcal{D}, \kappa)$  such that  $X$  is a subset-maximal envelope of  $\text{Coh}(\mathcal{D}, \kappa)$  and  $E \in X$ .

*Proof.* Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$  and  $\kappa$  a classifier.

► Let  $X \in \text{Coh}(\mathcal{D}, \kappa)$ ,  $x \in \mathcal{D}$ , and  $E \in X$ . Assume that  $E \subseteq x$ . From the second condition of the definition of an envelope,  $\exists y \in \mathcal{D}$  such that  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, y \rangle)$ . From coherence of  $X$ ,  $\kappa(x) = \kappa(y)$ , so  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ .

► From Property 2,  $\forall x \in \mathcal{D}$ ,  $x \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ , so  $\mathcal{D}$  satisfies the second condition of the definition of an envelope.  $\mathcal{D}$  also trivially satisfies the third condition of this same definition. From Property 4,  $\mathcal{D}$  is coherent, and hence satisfies the first condition of the same definition. So  $\mathcal{D} \in \text{Coh}(\mathcal{D}, \kappa)$ .

► Let  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ . We show that we can build a subset-maximal envelope  $X$  which contains  $E$ . From Property 4,  $\mathcal{D}$  is coherent. Then, we start with  $X_0 = \mathcal{D}$ . Assume that  $X_1 = X_0 \cup \{E\}$  is incoherent. Then,  $\exists y \in \mathcal{D}$  such that  $E \cup y$  is consistent and  $\exists z \in \mathcal{D}$  with  $E \subseteq z$  s.t.  $\kappa(y) \neq \kappa(z)$ . Since  $y$  is an instance, then  $E \subseteq y$ . From  $E \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ , it follows that

$\kappa(y) = \kappa(z) = \kappa(x)$ , which contradicts the incoherence of  $X_1$ . We then add to  $X_1$  all dwAXp's built from  $\mathcal{D}$  up to coherence.  $\square$

**Proposition 3.** *Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier.*

- $\mathcal{D} \subseteq \text{Irr}(\mathcal{D}, \kappa)$ .
- $\text{Irr}(\mathcal{D}, \kappa) \in \text{Coh}(\mathcal{D}, \kappa)$ .
- $\text{Irr}(\mathcal{D}, \kappa) = \bigcap_{X_i \in \mathcal{S}} X_i$ , where  $\mathcal{S}$  is the set of all subset-maximal envelopes of  $\text{Coh}(\mathcal{D}, \kappa)$ .

*Proof.* Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , and  $\kappa$  a classifier.

► Assume that  $x \in \mathcal{D}$  and  $x \notin \text{Irr}(\mathcal{D}, \kappa)$ . From Property 2,  $x \in \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ . Then, there exists a question  $\mathbf{Q}$  and  $\exists E \in \mathbf{L}_{\text{dw}}(\mathbf{Q})$  such that  $\{E, x\}$  is incoherent. So,  $E \cup x$  is consistent and  $\exists z \in \mathcal{D}$  such that  $E \subseteq z$  and  $\kappa(x) \neq \kappa(z)$ . From consistency of  $E \cup x$ , it follows that  $E \subseteq x$ ; which contradicts the fact that  $E$  is a dwAXp.

► From Def. 10,  $\text{Irr}(\mathcal{D}, \kappa) \subseteq \mathbb{E}(\mathbb{T})$  such that it is coherent and  $\text{Irr}(\mathcal{D}, \kappa) \subseteq \bigcup_{x \in \mathcal{D}} \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ . From the above result (i.e.,  $\mathcal{D} \subseteq \text{Irr}(\mathcal{D}, \kappa)$ ), it follows that  $\text{Irr}(\mathcal{D}, \kappa) \in \text{Coh}(\mathcal{D}, \kappa)$ .

► Let us show the third property. Let  $\mathbb{W} = \bigcup_{x \in \mathcal{D}} \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$  and  $\mathcal{S}$  be the set of all subset-maximal envelopes of  $\text{Coh}(\mathcal{D}, \kappa)$ . From Proposition 2,  $\mathcal{D}$  is an envelope. So  $\text{Coh}(\mathcal{D}, \kappa) \neq \emptyset$ . Let  $\mathcal{S} = \{X_1, \dots, X_k\}$  be the set of all subset-maximal envelopes. From Proposition 2,  $\forall i = 1, \dots, k$ ,  $X_i \subseteq \mathbb{W}$ .

Assume that  $E \in \text{Irr}(\mathcal{D}, \kappa)$ . Hence,  $E \in \mathbb{W}$  and  $\forall E' \in \mathbb{W}$ ,  $\{E, E'\}$  is coherent. From Coherence of every envelope  $X_i$ ,  $i = 1, \dots, k$ , it follows that  $X_i \cup \{E\}$  is coherent and from Maximality of  $X_i$ ,  $E \in X_i$ , thus  $E \in \bigcap_{X_i \in \mathcal{S}} X_i$ .

Assume now that  $E \in \bigcap_{X_i \in \mathcal{S}} X_i$  and  $E \notin \text{Irr}(\mathcal{D}, \kappa)$ . From the subset-maximality of  $\text{Irr}(\mathcal{D}, \kappa)$ ,  $\exists E' \in \mathbb{W}$  such that  $\{E, E'\}$  is incoherent. From Proposition 2,  $\exists X_j \in \mathcal{S}$  such that  $E' \in X_j$ . From the assumption  $E \in \bigcap_{X_i \in \mathcal{S}} X_i$ , it holds that  $E \in X_j$ , which contradicts the coherence of the envelope  $X_j$ .  $\square$

**Proposition 4.** *Let  $\mathbb{T}$  be a theory and  $\mathcal{D} \subseteq \mathbb{T}$ . For any classifier  $\kappa$ , there exists a decision-tree classifier  $\sigma$  on  $\mathbb{T}$  such that  $\forall y \in \mathcal{D}$ ,  $\sigma(y) = \kappa(y)$ .*

*Proof.* Let  $\mathcal{D}$  be a set of  $m$  distinct feature-vectors. A decision tree composed of at most  $m - 1$  nodes can always be built to distinguish between the  $m$  elements of  $\mathcal{D}$ , with each  $x \in \mathcal{D}$  corresponding to a leaf  $\ell_x$ . It then suffices to associate to each leaf  $\ell(x)$  the corresponding value  $\kappa(x)$ .  $\square$

**Proposition 5.** *Let  $\mathbb{T}$  be a theory,  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ ,  $\kappa$  a classifier and  $X$  an envelope under  $(\mathcal{D}, \kappa)$ . Then there exists a surrogate function  $\sigma_X : \mathbb{F} \rightarrow \mathbb{C}$  such that (1)  $\sigma_X$  and  $\kappa$  are equal on  $\mathcal{D}$  and (2) for all  $x \in \mathcal{D}$ ,  $E \in X$  and  $E \subseteq x$  implies  $E$  is a weak abductive explanation of  $\langle \mathbb{T}, \sigma_X, \mathbb{F}, x \rangle$ .*

*Proof.* It suffices to define  $\sigma_X$  and show that it has the desired properties. Let  $\mathbb{F}_X \subseteq \mathbb{F}$  be the instances  $z \in \mathbb{F}$  such that  $E \subseteq z$  for some  $E \in X$ . For  $z \in \mathbb{F}_X$ , define  $\sigma_X(z)$  to be equal to  $\kappa(x)$  if there exists  $E \in X$  and  $x \in \mathcal{D}$  such that  $E \subseteq z$  and  $E \subseteq x$ . By the definition of an envelope,  $E$  is a weak abductive explanation of  $\kappa$  on  $\mathcal{D}$  and so the value of  $\kappa(x)$  is identical for all  $x \in \mathcal{D}$  such that  $E \subseteq x$ . Furthermore, if  $E \subseteq z$  and  $E' \subseteq z$ , for  $E, E' \in X$ , and  $E \subseteq x$ ,  $E' \subseteq x'$ ,

where  $x, x' \in \mathcal{D}$ , then we must have  $\kappa(x) = \kappa(x')$  by coherence of  $X$ . Hence  $\sigma_X(z)$  is well-defined. For  $z \in \mathbb{F} \setminus \mathbb{F}_X$ , let  $\sigma_X(z)$  be equal to an arbitrary value from  $\mathbb{C}$ .

For  $x \in \mathcal{D}$ ,  $\sigma_X(x) = \kappa(x)$  since by definition of an envelope, there exists  $E \in X$  such that  $E \subseteq x$ . Furthermore, for all  $x \in \mathcal{D}$  and for all  $E \in X$  such that  $E \subseteq x$ ,  $E$  is a weak abductive explanation of  $\langle \mathbb{T}, \sigma, \mathbb{F}, x \rangle$ . since for all  $z$  such that  $E \subseteq z$ , we have  $\sigma_X(z) = \kappa(x)$  (by definition of  $\sigma_X$ ).  $\square$

## C Proofs of Theorems

**Theorem 1.** *The axioms of every set  $\mathbf{I}_{i=1,5}$  below are incompatible.*

- (I<sub>1</sub>) *Feasibility, Success, Coherence and Irreducibility.*
- (I<sub>2</sub>) *Feasibility, Coherence and Completeness.*
- (I<sub>3</sub>) *Strong Irreducibility and Strong Completeness.*
- (I<sub>4</sub>) *Feasibility, Validity, Success, Irreducibility and Monotonicity.*
- (I<sub>5</sub>) *Feasibility, Validity, Success, Irreducibility and CM.*

*Proof.*

► To show the incompatibility of the axioms in  $\mathbf{I}_1$  (Feasibility, Success, Coherence and Irreducibility), let us consider the theory and dataset  $\mathcal{D}_1$  below.

$\mathcal{D}_1$	$f_1$	$f_2$	$\kappa_1(x_i)$
$x_1$	0	1	0
$x_2$	1	0	1
$x_3$	0	0	2

Let us focus on the two instances  $x_1$  and  $x_2$ . The dwAXps of  $\mathbf{Q}_1 = \langle \mathbb{T}, \kappa_1, \mathcal{D}_1, x_1 \rangle$  and  $\mathbf{Q}_2 = \langle \mathbb{T}, \kappa_1, \mathcal{D}_1, x_2 \rangle$  are:

$$\begin{aligned} \mathbf{L}_{\text{dw}}(\mathbf{Q}_1) &= \{E_1, E_2\} & \mathbf{L}_{\text{dw}}(\mathbf{Q}_2) &= \{E'_1, E'_2\} \\ E_1 &= \{(f_2, 1)\} & E'_1 &= \{(f_1, 1)\} \\ E_2 &= \{(f_1, 0), (f_2, 1)\} & E'_2 &= \{(f_1, 1), (f_2, 0)\}. \end{aligned}$$

Let now  $\mathbf{L}$  be an explainer that satisfies Success, Feasibility, Coherence and Irreducibility.

From Success,  $\mathbf{L}(\mathbf{Q}_1) \neq \emptyset$  and  $\mathbf{L}(\mathbf{Q}_2) \neq \emptyset$ .

From Feasibility, Validity and Theorem 4,  $\mathbf{L}(\mathbf{Q}_1) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q}_1)$  and  $\mathbf{L}(\mathbf{Q}_2) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q}_2)$ .

From Irreducibility,  $\mathbf{L}(\mathbf{Q}_1) = \{E_1\}$  and  $\mathbf{L}(\mathbf{Q}_2) = \{E'_1\}$ . Note that  $E_1 \cup E'_1$  is consistent and  $\kappa(x_1) \neq \kappa(x_2)$ , which contradicts the fact that  $\mathbf{L}$  satisfies Coherence.

► To show the incompatibility of the axioms in  $\mathbf{I}_2$  (Feasibility, Coherence and Completeness), assume an explainer  $\mathbf{L}$  that satisfies the three axioms and consider the above theory, classifier and dataset  $\mathcal{D}_1$ . Proposition 1 ensures that  $\mathbf{L}$  satisfies Success (from Completeness) and Validity (from Success Feasibility and Coherence). Theorem 5 implies that  $\mathbf{L} = \mathbf{L}_{\text{dw}}$ , so  $\mathbf{L}(\mathbf{Q}_1) = \{E_1, E_2\}$  and  $\mathbf{L}(\mathbf{Q}_2) = \{E'_1, E'_2\}$ . Note that  $E_1 \cup E'_1$  is consistent while  $\kappa(x_1) \neq \kappa(x_2)$ , which contradicts Coherence of  $\mathbf{L}$ .

► To show the incompatibility of the axioms in  $\mathbf{I}_3$  (Strong Irreducibility and Strong Completeness), assume an explainer  $\mathbf{L}$  that satisfies them both and consider the classifier  $\kappa_1$ , dataset  $\mathcal{D}_1$  and question  $\mathbf{Q}_1 = \langle \mathbb{T}, \kappa_1, \mathcal{D}_1, x_1 \rangle$  (above). We complete the definition of  $\kappa_1$  over the whole feature space by setting  $\kappa_1(1, 1) = 0$ . Strong Completeness tells us that  $E_1, E_2 \in \mathbf{L}(\mathbf{Q}_1)$ , but this violates Strong Irreducibility since  $E_1 \subset E_2$ .

► To show the incompatibility of the axioms in  $\mathbf{I}_4$  (Feasibility, Validity, Success, Irreducibility, and Monotonicity), assume an explainer  $\mathbf{L}$  that satisfies them all, and consider the theory, classifier and dataset below.

$\mathcal{D}_1$	$f_1$	$f_2$	$\kappa_3(x_i)$
$x_1$	0	0	0
$x_2$	0	1	1

$$\begin{aligned} E_1 &= \emptyset \\ E_2 &= \{(f_1, 0)\} \\ E_3 &= \{(f_2, 0)\} \\ E_4 &= x_1 \end{aligned}$$

Consider now the question  $\mathbf{Q} = \langle \mathbb{T}, \kappa_3, \mathcal{D}_1, x_1 \rangle$ . From Feasibility,  $\mathbf{L}(\mathbf{Q}) \subseteq \{E_1, E_2, E_3, E_4\}$ . From Validity,  $E_1 \notin \mathbf{L}(\mathbf{Q})$  and  $E_2 \notin \mathbf{L}(\mathbf{Q})$ . From Irreducibility,  $E_4 \notin \mathbf{L}(\mathbf{Q})$ . From Success,  $\mathbf{L}(\mathbf{Q}) = \{E_3\}$ . Consider now the extended dataset  $\mathcal{D}_2$  below and the question  $\mathbf{Q}' = \langle \mathbb{T}, \kappa_3, \mathcal{D}_2, x_1 \rangle$ .

$\mathcal{D}_2$	$f_1$	$f_2$	$\kappa_3(x_i)$
$x_1$	0	0	0
$x_2$	0	1	1
$x_3$	1	0	1

From Monotonicity, since  $\mathcal{D}_1 \subset \mathcal{D}_2$ , then  $E_3 \in \mathbf{L}(\mathbf{Q}')$  while Validity ensures that  $E_3 \notin \mathbf{L}(\mathbf{Q}')$ .

► To show the incompatibility of the axioms in  $\mathbf{I}_5$  (Feasibility, Validity, Success, Irreducibility, and CM), it is sufficient to consider the above example. Note that  $\mathbf{L}(\mathbf{Q}) = \{E_3\}$  and  $\mathbf{L}(\mathbf{Q}') = \{E_4\}$  while  $\mathcal{D}_1 \subseteq \mathcal{D}_2$ .  $\square$

**Theorem 2.** *The axioms of every set  $\mathbf{C}_{i=1,5}$  below are compatible.*

- ( $\mathbf{C}_1$ ) Feasibility, Validity, Success, (Strong) Completeness, CM.
- ( $\mathbf{C}_2$ ) Feasibility, Validity, Success, (Strong) Irreducibility.
- ( $\mathbf{C}_3$ ) Feasibility, Validity, Success, Coherence, Monotonicity, CM, Strong Completeness.
- ( $\mathbf{C}_4$ ) Feasibility, Validity, Success, Coherence, Monotonicity, CM, Strong Irreducibility.
- ( $\mathbf{C}_5$ ) Feasibility, Validity, Coherence, (Strong) Irreducibility, Monotonicity, CM.

*Proof.*

► The axioms of the set  $\mathbf{C}_1$  (Feasibility, Validity, Success, (Strong) Completeness and Counter-Monotonicity) are compatible as  $\mathbf{L}_{\text{dw}}$  satisfies all of them (see Table 2).

► The axioms of the set  $\mathbf{C}_2$  (Feasibility, Validity, Success, (Strong) Irreducibility) are compatible since they are all satisfied by the explainer  $\mathbf{L}_{\text{dc}}$  (see Table 2).

► The axioms of the set  $\mathbf{C}_3$  (Feasibility, Validity, Success, Coherence, Monotonicity, CM, Strong Completeness) are compatible since the explainer  $\mathbf{L}_{\text{w}}$  satisfies all of them (see Table 2).

► The axioms of the set  $\mathbf{C}_4$  (Feasibility, Validity, Success, Coherence, Monotonicity, CM, Strong irreducibility) are compatible since the explainer  $\mathbf{L}_{\text{c}}$  satisfies all of them (see Table 2).

► The axioms of the set  $\mathbf{C}_5$  (Feasibility, Validity, Coherence, (Strong) Irreducibility, Monotonicity, CM) are compatible since the trivial explainer which returns the empty set for each question satisfies all of them.  $\square$

**Theorem 3.** *Considering all sets of axioms among those listed in Table 1 which include Feasibility and Validity, the only minimal incompatible sets are  $\mathbf{I}_1, \dots, \mathbf{I}_5$  and the only maximal compatible sets are  $\mathbf{C}_1, \dots, \mathbf{C}_5$ .*

*Proof.* Let  $\mathbf{C}$  be a compatible set of axioms satisfied by an explainer  $\mathbf{L}$ . The following five cases are mutually exclusive and cover all possible sets  $\mathbf{C}$ :

1. If  $\mathbf{C}$  contains Completeness, then by Theorem 5,  $\mathbf{L} = \mathbf{L}_{\text{dw}}$  and by Theorem 6,  $\mathbf{L}$  satisfies the axioms  $\mathbf{C}_1$ . It follows trivially that  $\mathbf{C}_1$  is a maximal compatible set of axioms.

2. If  $\mathbf{C}$  contains neither Completeness nor Success, then by Proposition 1 it cannot contain Strong Completeness. All axioms apart from these three (i.e.  $\mathbf{C}_5$ ) are satisfied by the explainer  $\mathbf{L}$  that returns the empty set for each question. It follows that  $\mathbf{C}_5$  is a maximal compatible set of axioms.
3. If  $\mathbf{C}$  contains Success and Irreducibility but not Completeness, then by Proposition 1,  $\mathbf{L}$  also satisfies Strong Irreducibility and hence the set of axioms  $\mathbf{C}_2$ . By the incompatibilities in Theorem 1,  $\mathbf{L}$  can satisfy none of Coherence, Strong Completeness, Monotonicity and CM. Hence  $\mathbf{C}_2$  is a maximal compatible set of axioms.
4. If  $\mathbf{C}$  contains Success and Strong Irreducibility but neither Completeness nor Irreducibility, then by Theorem 1 it cannot contain Strong Completeness. The explainer  $\mathbf{L}_{\text{c}}$  satisfies all other axioms (i.e.  $\mathbf{C}_4$ ). It follows that  $\mathbf{C}_4$  is a maximal compatible set of axioms.
5. The final case is when  $\mathbf{C}$  contains Success but neither Completeness nor (Strong) Irreducibility. All other axioms (i.e.  $\mathbf{C}_3$ ) are satisfied by the explainer  $\mathbf{L}_{\text{w}}$ . It follows that  $\mathbf{C}_3$  is a maximal compatible set of axioms.

The minimality of the incompatible sets  $\mathbf{I}_i$  ( $i = 1, \dots, 5$ ) is easily verified by checking that each proper subset of  $\mathbf{I}_i$  is a subset of one of the  $\mathbf{C}_j$  ( $j = 1, \dots, 5$ ) and hence is compatible. The above case analysis only required the incompatibility of  $\mathbf{I}_1, \dots, \mathbf{I}_5$  to cover all possible sets of axioms; it follows that there are no other minimal incompatible sets of axioms.  $\square$

**Theorem 4.** *An explainer  $\mathbf{L}$  satisfies Feasibility and Validity iff for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .*

*Proof.* Let  $\mathbf{L}$  be an explainer.

► Assume that  $\mathbf{L}$  that satisfies Feasibility and Validity. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question and  $E \in \mathbf{L}(\mathbf{Q})$ . Feasibility of  $\mathbf{L}$  implies  $E \subseteq x$ . Validity ensures that  $\forall y \in \mathcal{D}$  such that  $E \subseteq y$ , it holds that  $\kappa(x) = \kappa(y)$ . So,  $E$  is a dwAXp, i.e.,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .

► Assume that for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ . Then,  $\forall E \in \mathbf{L}(\mathbf{Q})$ , the following hold: i)  $E \subseteq x$ , so  $\mathbf{L}$  satisfies Feasibility, and ii)  $\forall y \in \mathcal{D}$ , if  $E \subseteq y$ , then  $\kappa(x) = \kappa(y)$ , so  $\mathbf{L}$  satisfies Validity.  $\square$

**Theorem 5.** *An explainer  $\mathbf{L}$  satisfies Feasibility, Validity and Completeness iff  $\mathbf{L} = \mathbf{L}_{\text{dw}}$ .*

*Proof.* Let  $\mathbf{L}$  be an explainer.

► Assume that  $\mathbf{L} = \mathbf{L}_{\text{dw}}$ . From Table 2,  $\mathbf{L}_{\text{dw}}$  satisfies Feasibility, Validity and Completeness.

► Assume now  $\mathbf{L}$  is an explainer that satisfies Feasibility, Validity and Completeness. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question.

- From Theorem 4,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .
- Let  $E \in \mathbf{L}_{\text{dw}}(\mathbf{Q})$  and assume  $E \notin \mathbf{L}(\mathbf{Q})$ . Completeness of  $\mathbf{L}$  implies that  $\exists y \in \mathcal{D}$  such that  $E \subseteq y$  and  $\kappa(x) \neq \kappa(y)$ , which contradicts the validity of  $\mathbf{L}_{\text{dw}}$ . So,  $\mathbf{L}_{\text{dw}}(\mathbf{Q}) \subseteq \mathbf{L}(\mathbf{Q})$ .

Thus,  $\mathbf{L}(\mathbf{Q}) = \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .  $\square$

**Theorem 6.** *The explainer  $\mathbf{L}_{\text{dw}}$  satisfies Feasibility, Validity, Success, (Strong) Completeness and CM. It violates the remaining axioms.*

*Proof.* Let us show the properties satisfied by the function  $\mathbf{L}_{\text{dw}}$ .

► Feasibility and Validity of  $\mathbf{L}_{\text{dw}}$  follow from its definition.

► *Success* of  $\mathbf{L}_{dw}$  follows from Property 2 (every instance  $x$  is a weak abductive explanation of  $\kappa(x)$ ).

► *Counter-Monotonicity*: Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle$  be two questions such that  $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathbb{F}(\mathbb{T})$ . Let  $E \in \mathbf{L}_{dw}(\mathbf{Q})$ . Hence,  $E \subseteq x$  and  $\forall y \in \mathcal{D}'$ , if  $E \subseteq y$ , then  $\kappa(y) = \kappa(x)$ . Since  $\mathcal{D} \subseteq \mathcal{D}'$ , we have  $\{y \in \mathcal{D} \mid E \subseteq y\} \subseteq \{y \in \mathcal{D}' \mid E \subseteq y\}$ . Hence  $E \in \mathbf{L}_{dw}(\mathbf{Q}')$ .

► *(Strong) Completeness*: Let us show that  $\mathbf{L}_{dw}$  satisfies Completeness. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $E \subseteq x$ . Assume that  $E \notin \mathbf{L}_{dw}(\mathbf{Q})$ . Then,  $E$  violates the validity condition in the definition of  $\mathbf{L}_{dw}$ , hence  $\exists y \in \mathcal{D}$  such that  $E \subseteq y$  and  $\kappa(y) \neq \kappa(x)$ . So,  $\mathbf{L}$  satisfies Completeness. Strong Completeness follows from Completeness by Proposition 1.

► *Coherence*: From Theorem 1, since  $\mathbf{L}_{dw}$  satisfies Feasibility and Completeness, we can deduce that Coherence is violated (the set  $\mathbf{I}_2$  being incompatible).

► *Irreducibility*: From Theorem 1, since  $\mathbf{L}_{dw}$  satisfies Feasibility and Completeness, it violates Irreducibility (the set  $\mathbf{I}_3$  being incompatible).

► *Monotonicity*: Consider the above example using the classifier  $\kappa_3$ . Note that  $\mathbf{L}_{dw}(\mathbf{Q}) = \{E_3\}$  and  $\mathbf{L}_{dw}(\mathbf{Q}') = \{E_4\}$  while  $\mathcal{D}_1 \subseteq \mathcal{D}_2$ .  $\square$

**Theorem 7.** *An explainer  $\mathbf{L}$  satisfies Feasibility, Validity, Monotonicity and Strong Completeness iff  $\mathbf{L} = \mathbf{L}_w$ .*

*Proof.* Let  $\mathbf{L}$  be an explainer.

► Assume that  $\mathbf{L}$  satisfies Feasibility, Validity, Monotonicity and Strong Completeness. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathbb{F}(\mathbb{T}), x \rangle$ . Since  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ , then from Monotonicity of  $\mathbf{L}$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}(\mathbf{Q}')$ . From Feasibility and Validity, it follows from Theorem 4 that  $\mathbf{L}(\mathbf{Q}') \subseteq \mathbf{L}_{dw}(\mathbf{Q}')$ . By definition of  $\mathbf{L}_w$ ,  $\mathbf{L}_w(\mathbf{Q}) = \mathbf{L}_{dw}(\mathbf{Q}')$ . So,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_w(\mathbf{Q})$ .

Consider now some  $E \in \mathbf{L}_w(\mathbf{Q})$  and  $E \notin \mathbf{L}(\mathbf{Q})$ . Then,  $E \subseteq x$ . From Strong Completeness,  $\exists y \in \mathbb{F}(\mathbb{T})$  such that  $E \subseteq y$  and  $\kappa(x) \neq \kappa(y)$ . This contradicts the validity of  $E$  on  $\mathbb{F}(\mathbb{T})$ . So,  $\mathbf{L}_w(\mathbf{Q}) \subseteq \mathbf{L}(\mathbf{Q})$ .

► Theorem 8 shows that  $\mathbf{L}_w$  satisfies the four axioms.  $\square$

**Theorem 8.** *The explainer  $\mathbf{L}_w$  satisfies Feasibility, Validity, Success, Strong Completeness, Coherence, Monotonicity and CM. It violates (Strong) Irreducibility and Completeness.*

*Proof.*

► *Feasibility and Validity* follow from the definition.

► *Success* follows from the fact that for any  $x \in \mathbb{F}(\mathbb{T})$ ,  $x \in \mathbf{L}_w(\mathbf{Q})$  where  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathbb{F}(\mathbb{T}), x \rangle$ .

► *Coherence*: Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}, y \rangle$  such that  $\kappa(x) \neq \kappa(y)$ . Assume that  $E \in \mathbf{L}_w(\mathbf{Q})$ ,  $E' \in \mathbf{L}_w(\mathbf{Q}')$  and  $E \cup E'$  is consistent. Then,  $\exists z \in \mathbb{F}(\mathbb{T})$  such that  $E \cup E' \subseteq z$ . From Validity of  $\mathbf{L}_w$ ,  $\kappa(x) = \kappa(z)$  and  $\kappa(y) = \kappa(z)$ , hence  $\kappa(x) = \kappa(y)$ , which contradicts the assumption.

► *Monotonicity and Counter-Monotonicity* follow from the fact that for all  $\mathcal{D} \subseteq \mathcal{D}'$ , for all  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  and  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}', x \rangle$ ,  $\mathbf{L}_w(\mathbf{Q}) = \mathbf{L}_w(\mathbf{Q}')$ .

► *Strong Completeness*: Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question and  $E \subseteq x$ . Assume  $E \notin \mathbf{L}_w(\mathbf{Q})$ . By definition of  $\mathbf{L}_w$ ,  $\exists y \in \mathbb{F}(\mathbb{T})$  such that  $E \subseteq y$  and  $\kappa(x) \neq \kappa(y)$ . Hence,  $\mathbf{L}_w$  satisfies Strong Completeness.

► *Irreducibility and Completeness* are violated since from Theorem 1, Completeness is incompatible with the pair (Feasibility, Coherence) and Irreducibility is incompatible with the tuple (Feasibil-

ity, Success, Coherence). Strong Irreducibility is violated since from Proposition 1, it is incompatible with Strong Completeness.  $\square$

**Theorem 9.** *An explainer  $\mathbf{L}$  satisfies Feasibility, Validity, and Irreducibility iff for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{dc}(\mathbf{Q})$ .*

*Proof.* Let  $\mathbf{L}$  be an explainer.

► Assume now  $\mathbf{L}$  is an explainer that satisfies Feasibility, Validity, and Irreducibility. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question. Assume  $E \in \mathbf{L}(\mathbf{Q})$ . From Theorem 4,  $E \in \mathbf{L}_{dw}(\mathbf{Q})$ , so  $E$  is a weak abductive explanation. Assume now that  $E \notin \mathbf{L}_{dc}(\mathbf{Q})$ , so  $E$  is not subset-minimal. Then,  $\exists E' \in \mathbf{L}_{dw}(\mathbf{Q})$  such that  $E' \subset E$ . hence,  $\exists l \in E \setminus E'$  and  $\exists y \in \mathcal{D}$  such that  $E \setminus \{l\} \subseteq y$  and  $\kappa(x) = \kappa(y)$ , which contradicts Irreducibility of  $\mathbf{L}$ . Hence,  $E \in \mathbf{L}_{dc}(\mathbf{Q})$ .

► Assume that for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{dc}(\mathbf{Q})$ . From Property 2,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{dw}(\mathbf{Q})$ . From Theorem 4,  $\mathbf{L}$  satisfies Feasibility and Validity. From Table 2,  $\mathbf{L}_{dc}$  satisfies Irreducibility, hence so does  $\mathbf{L}$ .  $\square$

**Theorem 10.** *The explainer  $\mathbf{L}_{dc}$  satisfies Feasibility, Validity, Success and (Strong) Irreducibility. It violates the remaining axioms.*

*Proof.* Let us show the properties satisfied by  $\mathbf{L}_{dc}$ .

► *Feasibility and Validity* of  $\mathbf{L}_{dc}$  follow from the definition of the function.

► *Success* of  $\mathbf{L}_{dc}$  follows from Property 2 (every instance  $x$  has a  $\mathbf{L}_{dc}$ , namely itself, and hence must have a subset-minimal  $\mathbf{L}_{dc}$ ).

► *Irreducibility*: Irreducibility of  $\mathbf{L}_{dc}$  follows from its minimality condition.

► *Coherence, Completeness, Monotonicity and CM* are violated due to their incompatibility with Irreducibility (see Theorem 1).  $\square$

**Theorem 11.** *The function  $\mathbf{L}_c$  satisfies all the axioms except Irreducibility and (Strong) Completeness.*

*Proof.* Let us show the properties satisfied by  $\mathbf{L}_c$ .

► *Success, Strong Irreducibility, Coherence* of  $\mathbf{L}_c$  have been shown in [14].

► *Feasibility and Validity* of  $\mathbf{L}_c$  follow from the definition of the function.

► *Monotonicity and CM* follows straightforwardly from the fact that for any question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}_c(\mathbf{Q}) = \mathbf{L}_{dc}(\mathbf{Q}')$  where  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathbb{F}(\mathbb{T}), x \rangle$ .

► *Irreducibility and Completeness*: To show their violation by  $\mathbf{L}_c$ , consider the following theory, classifier, the dataset consisting of  $x_1, x_2$  and the question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x_1 \rangle$ .

$\mathbb{F}(\mathbb{T})$	$f_1$	$f_2$	$\kappa(x_i)$
$x_1$	0	0	0
$x_2$	0	1	1
$x_3$	1	0	1
$x_4$	1	1	0

Note that  $\mathbf{L}_c(\mathbf{Q}) = \{\{(f_1, 0), (f_2, 0)\}\}$  and  $\nexists y \in \mathcal{D}$  such that  $(f_2, 0) \in y$  and  $\kappa(y) \neq \kappa(x_1)$ . Hence,  $\{\{(f_1, 0), (f_2, 0)\}\}$  can be reduced to  $\{(f_2, 0)\}$ . The same example shows that Completeness is violated (indeed,  $\{(f_2, 0)\} \notin \mathbf{L}_c(\mathbf{Q})$ ).

► *Strong Completeness*: Consider the following theory, classifier, the dataset made of  $x_1, x_2$  and the question  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x_1 \rangle$ .

$\mathbb{F}(\mathbb{T})$	$f_1$	$f_2$	$\kappa(x_i)$
$x_1$	0	0	0
$x_2$	0	1	0
$x_3$	1	0	1
$x_4$	1	1	1

Note that  $\mathbf{L}_c(\mathbf{Q}) = \{(f_1, 0)\}$ . So,  $x_1 \notin \mathbf{L}_c(\mathbf{Q})$  while  $\exists y \in \mathbb{F}(\mathbb{T})$  such that  $x_1 \subseteq y$  and  $\kappa(y) \neq \kappa(x_1)$ .  $\square$

**Theorem 12.** *If an explainer  $\mathbf{L}$  satisfies Feasibility, Success and Coherence, then for any question  $\mathbf{Q}$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ .*

*Proof.* Let  $\mathbf{L}$  be an explainer that satisfies Feasibility, Success and Coherence. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question and  $E \in \mathbf{L}(\mathbf{Q})$ . Feasibility leads to  $E \subseteq x$ . Let  $Y = \{y \in \mathcal{D} \mid E \subseteq y\}$ . Assume that  $z \in Y$  and  $\kappa(z) \neq \kappa(x)$ . Let  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}, z \rangle$  be a question. Success implies that  $\mathbf{L}(\mathbf{Q}') \neq \emptyset$ . So,  $\exists E' \in \mathbf{L}(\mathbf{Q}')$ . Feasibility ensures  $E' \subseteq z$ . Hence,  $E \cup E' \subseteq z$ , which means  $E \cup E'$  is consistent. This contradicts Coherence which guarantees the inconsistency of  $E \cup E'$ .  $\square$

**Theorem 13.** *An explainer  $\mathbf{L}$  satisfies Feasibility, Success and Coherence iff  $\mathbf{L}$  is a coherent explainer.*

*Proof.* Assume that an explainer  $\mathbf{L}$  satisfies Feasibility, Success and Coherence. From Theorem 12,  $\forall \mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{L}(\mathbf{Q}) \subseteq \mathbf{L}_{\text{dw}}(\mathbf{Q})$ . From Coherence of  $\mathbf{L}$ , the set  $\bigcup_{x \in \mathcal{D}} \mathbf{L}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$  is coherent. From Success,  $\forall x \in \mathcal{D}$ ,  $\mathbf{L}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle) \neq \emptyset$ . Hence,  $\bigcup_{x \in \mathcal{D}} \mathbf{L}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle) \in \text{Coh}(\mathcal{D}, \kappa)$  and so  $\mathbf{L}$  is a coherent explainer.

Let us now show that  $\mathbf{L}_{\text{co}}$  satisfies the three axioms. Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question and  $X = \bigcup_{x \in \mathcal{D}} \mathbf{L}_{\text{co}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle)$ . Since  $X \in \text{Coh}(\mathcal{D}, \kappa)$ , then it is coherent, so  $\mathbf{L}_{\text{co}}$  satisfies Coherence. From Definition 7,  $\exists E \in X$  such that  $E \subseteq x$ . So,  $\mathbf{L}_{\text{co}}$  satisfies Success. Feasibility is satisfied since by definition, every explanation is a subset of the instance.  $\square$

**Theorem 14.** *Any coherent explainer  $\mathbf{L}_{\text{co}}$  satisfies Feasibility, Validity, Success, Coherence. It violates Irreducibility and Completeness.*

*Proof.* Theorem 13 shows that  $\mathbf{L}_{\text{co}}$  satisfies Feasibility, Success, Coherence. From Proposition 1, Validity follows from Feasibility, Success, Coherence. Theorem 1 shows that Coherence is incompatible with Irreducibility (set  $\mathbf{I}_1$ ) and Completeness (set  $\mathbf{I}_2$ ).  $\square$

**Theorem 15.** *The function  $\mathbf{L}_{\text{tr}}$  satisfies all the axioms except (Strong) Irreducibility and (Strong) Completeness.*

*Proof.*

*Success, Feasibility, Coherence* follow from Theorem 13 ( $\mathbf{L}_{\text{tr}}$  is a coherent explainer). Validity follows from Proposition 1.

*Monotonicity and CM* follow straightforwardly from the definition of  $\mathbf{L}_{\text{tr}}$  which provides a single explanation (the instance itself) for every question.

*Irreducibility and Completeness* are violated due to Theorem 1.

*Strong Irreducibility and Strong Completeness:* Consider the theory and classifier below.

$\mathbb{F}(\mathbb{T})$	$f_1$	$f_2$	$\kappa(x_i)$
$x_1$	0	0	0
$x_2$	0	1	1
$x_3$	1	0	0
$x_4$	1	1	1

Consider the dataset  $\mathcal{D} = \{x_1, x_2\}$ . Note that  $\mathbf{L}_{\text{tr}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x_1 \rangle) = \{x_1\}$ . So,  $\{(f_2, 0)\} \notin \mathbf{L}_{\text{tr}}(\langle \mathbb{T}, \kappa, \mathcal{D}, x_1 \rangle)$  while  $\exists y \in \mathbb{F}(\mathbb{T})$  such that  $(f_2, 0) \in y$  and  $\kappa(x_1) \neq \kappa(y)$ . Thus, Strong Completeness is violated. From the same example, note that for  $l = (f_1, 0)$ ,  $x_1 \setminus \{l\} = \{(f_2, 0)\}$  while  $\exists y \in \mathbb{F}(\mathbb{T})$  such that  $(f_2, 0) \in y$  and  $\kappa(x_1) \neq \kappa(y)$ . Thus, Strong Irreducibility is violated.  $\square$

**Theorem 16.** *The function  $\mathbf{L}_{\text{ir}}$  satisfies Feasibility, Validity, Success and Coherence. It violates all the other axioms.*

*Proof.*

*Success, Feasibility, Validity and Coherence* follow from Property 7 and Theorem 13.

*Monotonicity:* Consider the sample  $\mathcal{D}_1$  below.

$\mathcal{D}_1$	$f_1$	$f_2$	$\kappa(x_i)$
$x_1$	0	0	0
$x_2$	1	0	1

From  $\mathcal{D}_1$ , two dwAXp ( $E_i$ ) are generated for  $x_1$  and two others ( $E'_j$ ) concern  $x_2$ .

$$\begin{aligned} E_1 &= \{(f_1, 0)\} & E'_1 &= \{(f_1, 1)\} \\ E_2 &= \{(f_1, 0), (f_2, 0)\} & E'_2 &= \{(f_1, 1), (f_2, 0)\} \end{aligned}$$

Note that the four sets are not conflicting. Thus,  $\text{Irr}(\mathcal{D}_1, \kappa) = \{E_1, E_2, E'_1, E'_2\}$  and  $\mathbf{L}_{\text{ir}}(\langle \mathbb{T}, \kappa, \mathcal{D}_1, x_1 \rangle) = \{E_1, E_2\}$ .

Consider now the extended sample  $\mathcal{D}_2 = \mathcal{D}_1 \cup \{x_3\}$  where  $x_3$  is as shown below.

	$f_1$	$f_2$	$\kappa(x_3)$
$x_3$	1	1	1

Note that  $\mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_1 \rangle) = \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_1, x_1 \rangle)$ ,  $\mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_2 \rangle) = \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_1, x_2 \rangle)$  and  $\mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_3 \rangle) = \{E'_1, \{(f_2, 1)\}, x_3\}$ . Note that  $\{E_1, \{(f_2, 1)\}\}$  is incoherent, then  $E_1 \notin \text{Irr}(\mathcal{D}_2, \kappa)$ , and so  $E_1 \notin \mathbf{L}_{\text{ir}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_1 \rangle)$ .

*Counter-Monotonicity:* To show that  $\mathbf{L}_{\text{ir}}$  violates Counter-Monotonicity, consider the sample  $\mathcal{D}_1$  below.

$\mathcal{D}_1$	$f_1$	$f_2$	$\kappa(x_i)$
$x_1$	1	0	0
$x_2$	0	1	1

From  $\mathcal{D}_1$ , three dwAXp ( $E_i$ ) are generated for  $x_1$  and three others ( $E'_j$ ) concern  $x_2$ .

$$\begin{aligned} E_1 &= \{(f_1, 1)\} & E'_1 &= \{(f_1, 0)\} \\ E_2 &= \{(f_2, 0)\} & E'_2 &= \{(f_2, 1)\} \\ E_3 &= \{(f_1, 1), (f_2, 0)\} & E'_3 &= \{(f_1, 0), (f_2, 1)\} \end{aligned}$$

Note that  $E_1$  and  $E'_2$  are conflicting. The same holds for the pair  $E'_1$  and  $E_2$ . Hence,  $\mathbf{L}_{\text{ir}}(\langle \mathbb{T}, \kappa, \mathcal{D}_1, x_1 \rangle) = \{E_3\}$ . Consider now the extended sample  $\mathcal{D}_2 = \mathcal{D}_1 \cup \{x_3\}$  where  $x_3$  is as shown below.

	$f_1$	$f_2$	$\kappa(x_3)$
$x_3$	0	0	0

Note that  $\mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_1 \rangle) = \mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_1, x_1 \rangle)$  while  $\mathbf{L}_{\text{dw}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_2 \rangle) = \{E'_2, E'_3\}$ . Then,  $\mathbf{L}_{\text{ir}}(\langle \mathbb{T}, \kappa, \mathcal{D}_2, x_1 \rangle) = \{E_2, E_3\}$ .

*Irreducibility and Completeness* are violated due to their incompatibility with Coherence (sets  $\mathbf{I}_1$  and  $\mathbf{I}_3$ ) respectively in Theorem 1).

*Strong Irreducibility and Strong Completeness:* Consider the counter-example given for Counter-Monotonicity, namely the

dataset  $\mathcal{D}_2$  and assume that  $\kappa(x_4) = 0$ . Note that the set  $\{\{(f_1, 1)\}, \{(f_2, 1)\}\}$  is incoherent under  $(\mathcal{D}_2, \kappa)$ . Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}_2, x_1 \rangle$ . So,  $\mathbf{L}_{\text{ir}}(\mathbf{Q}) = \{E_2, E_3\}$ . Note that  $E_1 \notin \mathbf{L}_{\text{ir}}(\mathbf{Q})$  while  $\nexists y \in \mathbb{F}(\mathbb{T})$  such that  $E_1 \subseteq y$  and  $\kappa(y) \neq \kappa(x_3)$ . So, Strong Completeness is violated by  $\mathbf{L}_{\text{ir}}$ .

Consider now the question  $\mathbf{Q}' = \langle \mathbb{T}, \kappa, \mathcal{D}_2, x_3 \rangle$ . Note that  $\mathbf{L}_{\text{ir}}(\mathbf{Q}') = \{x_3\}$ . Let  $l = (f_1, 0)$ . Note that  $\forall y \in \mathbb{F}(\mathbb{T})$  such that  $x_3 \setminus \{l\} \in y$ ,  $\kappa(y) = \kappa(x_3)$ , so Strong Irreducibility is violated.  $\square$

**Theorem 17.** *Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$  be a question, where  $\kappa$  can be evaluated in polynomial time. Testing whether  $E \in \mathbf{L}_{\text{ir}}(\mathbf{Q})$  can be achieved in polynomial time.*

*Proof.* Testing whether  $E$  is a weak abductive explanation can be achieved in time  $O(mn)$ , where  $n$  is the number of features and  $m$  the number of instances in the dataset  $D$  [14]. Testing whether a weak abductive explanation  $E$  is irrefutable amounts to checking that no counter-example exists. A counter-example is another instance  $x' \in \mathcal{D}$  giving a distinct prediction (i.e.  $\kappa(x') \neq \kappa(x)$ ) together with a weak abductive explanation  $E'$  of  $x'$  such that  $E$  and  $E'$  attack each other. We claim that, for a given  $x' \in \mathcal{D}$ , we only need to test  $E' = x'[F \setminus (ft(E) \cap \{i \mid x_i \neq x'_i\})]$  as a putative weak abductive explanation, where  $ft(E)$  is the set of features assigned a value in  $E$  and  $x_i$  is the value of feature  $i$  in vector  $x$ .

Clearly if this  $E'$  (along with  $x'$ ) provides a counter-example, then  $E$  is not irrefutable. For a given  $x'$ , suppose that  $E' = x'[F \setminus (ft(E) \cap \{i \mid x_i \neq x'_i\})]$  is not a counter-example, but there is another  $E''$  which is (for  $x'$ ). Since  $E''$  is a counter-example (along with  $x'$ ),  $E''$  is a weak abductive explanation of  $\kappa(x') = c'$ . However, by supposition,  $x'[F \setminus (ft(E) \cap \{i \mid x_i \neq x'_i\})]$  is not a weak abductive explanation of the decision  $\kappa(x') = c'$ , so there exists  $u \in \mathcal{D}$  such that  $\kappa(u) \neq \kappa(x')$  and  $u[ft(E')] = x'[ft(E')]$ . It follows that  $ft(E'')$  cannot be a subset of  $ft(E')$ . Let  $j$  be a feature in  $ft(E'')$  but not in  $ft(E')$ : hence  $j$  must belong to  $ft(E) \cap \{i \mid x_i \neq x'_i\}$ . However, we know that  $E$  and  $E''$  attack each other, so they must agree on their intersection, which is contradicted by the existence of  $j$ . Hence, we have a proof of the claim, by contradiction.

The number of  $x' \in \mathcal{D}$  to be tested is bounded by  $m$ , the size of the dataset  $\mathcal{D}$ . For each such  $x'$ , the calculation of  $E' = x'[F \setminus (ft(E) \cap \{i \mid x_i \neq x'_i\})]$  is  $O(n)$ , where  $n$  is the number of features, and testing whether this  $E'$  is a weak abductive explanation can be achieved in  $O(mn)$  time. Hence, total complexity is  $O(n^2 m^2)$ .  $\square$

**Theorem 18.** *A surrogate explainer  $\mathbf{L}_{\text{DT}}$  can be found in polynomial time.*

*Proof.* This follows from the fact that standard algorithms, such as ID3, build a decision tree in time which is polynomial in the size of the resulting tree and the size  $m$  of the dataset. We have seen in the proof of Proposition 4 that the number of nodes in the decision tree is bounded by  $m$ .  $\square$

**Theorem 19.** *The explainer  $\mathbf{L}_{\text{su}}$  satisfies Feasibility, Validity, Success and Coherence. It violates the remaining axioms.*

*Proof.* Let  $\mathbf{Q} = \langle \mathbb{T}, \kappa, \mathcal{D}, x \rangle$ ,  $\mathbf{Q}' = \langle \mathbb{T}, \sigma, \mathbb{F}(\mathbb{T}), x \rangle$  and  $\sigma$  be a surrogate function on  $(\mathbb{T}, \kappa, \mathcal{D})$ . Recall that  $x \in \mathcal{D}$ .

From [14],  $\mathbf{L}_{\text{dw}}$  satisfies success when defined on the feature space. Hence,  $\mathbf{L}_{\text{su}}(\mathbf{Q}) \neq \emptyset$  and so it satisfies Success.

Let  $E \in \mathbf{L}_{\text{su}}(\mathbf{Q})$ , then  $E \in \mathbf{L}_{\text{dw}}(\mathbf{Q}')$ . By definition of a weak abductive explanation,  $E \subseteq x$ , so  $\mathbf{L}_{\text{su}}$  satisfies Feasibility. Furthermore,  $\forall z \in \mathbb{F}(\mathbb{T})$  such that  $E \subseteq z$ ,  $\sigma(z) = \sigma(x)$ . Since  $\mathcal{D} \subseteq \mathbb{F}(\mathbb{T})$ ,

then  $\forall z \in \mathcal{D}$ ,  $\sigma(z) = \sigma(x)$ . Since  $\sigma$  is a surrogate function on  $(\mathbb{T}, \kappa, \mathcal{D})$ , we have  $\sigma(z) = \sigma(x) = \kappa(z) = \kappa(x)$ . So  $\mathbf{L}_{\text{su}}$  satisfies Validity.

Coherence of  $\mathbf{L}_{\text{su}}$  follows from the Coherence of  $\mathbf{L}_{\text{dw}}$  when defined on the feature space [14].

Violation of Irreducibility follows from violation of the property by  $\mathbf{L}_{\text{dw}}$ .

To show that  $\mathbf{L}_{\text{su}}$  violates Monotonicity and CM, consider the theory  $\mathbb{T}$  below, a classifier  $\kappa$ , its surrogate models  $\sigma_1, \sigma_2$ , and two datasets  $\mathcal{D}_1 = \{x_1, x_2\}$  and  $\mathcal{D}_2 = \{x_1, x_2, x_3\}$  on which the surrogate models have been trained respectively.

$\mathbb{F}(\mathbb{T})$	$f_1$	$f_2$	$\kappa(x_i)$	$\sigma_1(x_i)$	$\sigma_2(x_i)$
$x_1$	0	0	0	0	0
$x_2$	0	1	1	1	1
$x_3$	1	0	0	1	0
$x_4$	1	1	1	1	0

Let  $\mathbf{Q}_1 = \langle \mathbb{T}, \kappa, \mathcal{D}_1, x_1 \rangle$ ,  $\mathbf{Q}'_1 = \langle \mathbb{T}, \kappa, \mathcal{D}_2, x_1 \rangle$ ,  $\mathbf{Q}_2 = \langle \mathbb{T}, \kappa, \mathcal{D}_1, x_2 \rangle$ ,  $\mathbf{Q}'_2 = \langle \mathbb{T}, \kappa, \mathcal{D}_2, x_2 \rangle$  be questions. Note that  $\mathbf{L}_{\text{su}}(\mathbf{Q}_1) = \{x_1\}$  while  $\mathbf{L}_{\text{su}}(\mathbf{Q}'_1) = \{x_1, \{(f_2, 0)\}\}$ , hence CM is violated. Note also that  $\mathbf{L}_{\text{su}}(\mathbf{Q}_2) = \{x_2, \{(f_2, 1)\}\}$  while  $\mathbf{L}_{\text{su}}(\mathbf{Q}'_2) = \{x_2\}$ , hence Monotonicity is violated.

Completeness and Strong Completeness are violated since  $\{(f_2, 0)\} \notin \mathbf{L}_{\text{su}}(\mathbf{Q}_1)$  while  $\nexists y \in \mathcal{D}_1$  (resp.  $\nexists y \in \mathbb{F}(\mathbb{T})$ ) such that  $\{(f_2, 0)\} \subseteq y$  and  $\kappa(y) \neq \kappa(x_1)$ .  $\square$

## References

- [1] L. Amgoud. Explaining black-box classifiers: Properties and functions. *International Journal of Approximate Reasoning*, 155:40–65, 2023.
- [2] L. Amgoud and J. Ben-Naim. Axiomatic foundations of explainability. In *IJCAI*, pages 636–642, 2022.
- [3] L. Amgoud, P. Muller, and H. Trenquier. Leveraging argumentation for generating robust sample-based explanations. In *IJCAI*, pages 3104–3111, 2023. URL <https://doi.org/10.24963/ijcai.2023/346>.
- [4] G. Audemard, F. Koriche, and P. Marquis. On tractable XAI queries based on compiled representations. In *KR*, pages 838–849, 2020. doi: 10.24963/KR.2020/86. URL <https://doi.org/10.24963/kr.2020/86>.
- [5] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J. Lagniez, and P. Marquis. Trading complexity for sparsity in random forest explanations. In *AAAI*, pages 5461–5469. AAAI Press, 2022.
- [6] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis. On preferred abductive explanations for decision trees and random forests. In *IJCAI*, pages 643–650, 2022.
- [7] G. Audemard, S. Bellart, J. Lagniez, and P. Marquis. Computing abductive explanations for boosted regression trees. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*, pages 3432–3441. ijcai.org, 2023.
- [8] G. Audemard, J. Lagniez, P. Marquis, and N. Szczepanski. Computing abductive explanations for boosted trees. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4699–4711, 2023.
- [9] R. Boumazouza, F. C. Alili, B. Mazure, and K. Tabia. ASTERYX: A model-Agnostic SaT-basEd appRoach for sYmbolic and score-based eXplanations. In *CIKM '21*, pages 120–129, 2021.
- [10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [11] R. Byrne. Semifactual “even if” thinking. *Thinking and reasoning*, 8 (1):41–67, 2002.
- [12] C. Carbonnel, M. C. Cooper, and J. Marques-Silva. Tractable explaining of multivariate decision trees. In *KR*, pages 127–135, 2023.
- [13] Z. Chen and S. Toda. The complexity of selecting maximal solutions. *Inf. Comput.*, 119(2):231–239, 1995.
- [14] M. C. Cooper and L. Amgoud. Abductive explanations of classifiers under constraints: Complexity and properties. In *ECAI*, volume 372, pages 469–476. IOS Press, 2023.
- [15] M. C. Cooper and J. Marques-Silva. On the tractability of explaining decisions of classifiers. In L. D. Michel, editor, *CP 2021*, volume 210 of *LIPICs*, pages 21:1–21:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

- [16] M. C. Cooper and J. Marques-Silva. Tractability of explaining classifier decisions. *Artificial Intelligence*, 316:103841, 2023. doi: <https://doi.org/10.1016/j.artint.2022.103841>.
- [17] A. Darwiche and A. Hirth. On the reasons behind decisions. In *24th European Conference on Artificial Intelligence ECAI*, volume 325, pages 712–720. IOS Press, 2020.
- [18] A. de Colnet and P. Marquis. On the complexity of enumerating prime implicants from decision-DNNF circuits. In L. D. Raedt, editor, *IJCAI*, pages 2583–2590, 2022.
- [19] A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 590–601, 2018.
- [20] European Commission. Regulatory framework proposal on artificial intelligence, 2021. URL <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.
- [21] J. Ferreira, M. de Sousa Ribeiro, R. Gonçalves, and J. Leite. Looking inside the black-box: Logic-based explanations for neural networks. In *19th International Conference on Principles of Knowledge Representation and Reasoning, KR*, page In press, 2022.
- [22] X. Huang, Y. Izza, A. Ignatiev, M. C. Cooper, N. Asher, and J. Marques-Silva. Tractable explanations for d-DNNF classifiers. In *AAAI*, pages 5719–5728. AAAI Press, 2022. doi: 10.1609/AAAI.V36I5.20514. URL <https://doi.org/10.1609/aaai.v36i5.20514>.
- [23] A. Ignatiev and J. P. M. Silva. SAT-based rigorous explanations for decision lists. In C. Li and F. Manyà, editors, *SAT*, volume 12831, pages 251–269. Springer, 2021.
- [24] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI 2019*, pages 1511–1519. AAAI Press, 2019.
- [25] A. Ignatiev, N. Narodytska, and J. Marques-Silva. On relating explanations and adversarial examples. In *NeurIPS*, pages 15857–15867, 2019.
- [26] Y. Izza and J. Marques-Silva. On explaining random forests with SAT. In Z. Zhou, editor, *IJCAI*, pages 2584–2591. ijcai.org, 2021.
- [27] Y. Izza, A. Ignatiev, and J. Marques-Silva. On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75:261–321, 2022.
- [28] Y. Izza, A. Ignatiev, P. J. Stuckey, and J. Marques-Silva. Delivering inflated explanations. In *Thirty-Eighth Conference on Artificial Intelligence, AAAI*, pages 12744–12753. AAAI Press, 2024.
- [29] R. Luss, P. Chen, A. Dhurandhar, P. Sattigeri, K. Shanmugam, and C. Tu. Generating contrastive explanations with monotonic attribute functions. *CoRR*, 2019. URL <http://arxiv.org/abs/1905.12698>.
- [30] J. Marques-Silva, T. Gerspacher, M. C. Cooper, A. Ignatiev, and N. Narodytska. Explanations for monotonic classifiers. In M. Meila and T. Zhang, editors, *ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 7469–7479. PMLR, 2021.
- [31] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986.
- [32] J. R. Quinlan. Simplifying decision trees. *Int. J. Man Mach. Stud.*, 27(3):221–234, 1987.
- [33] M. T. Ribeiro, S. Singh, and C. 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*, page 1135–1144, 2016.
- [34] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *(AAAI-18)*, pages 1527–1535, 2018.
- [35] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. GLocalX - from local to global explanations of black box AI models. *Artificial Intelligence*, 294:103457, 2021.
- [36] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining Bayesian network classifiers. In *IJCAI*, pages 5103–5111, 2018.
- [37] The White House OSTP. Blueprint for an AI Bill of Rights : Making Automated Systems Work for the American People, 2022. URL <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>.
- [38] UCI Machine Learning Repository. Zoo, 2017. URL <https://www.kaggle.com/datasets/uciml/zoo-animal-classification>.
- [39] P. E. Utgoff. Incremental induction of decision trees. *Mach. Learn.*, 4: 161–186, 1989.
- [40] I. van der Linden, H. Haned, and E. Kanoulas. Global aggregations of local explanations for black box models. *CoRR*, abs/1907.03039, 2019. URL <http://arxiv.org/abs/1907.03039>.
- [41] S. Wachter, B. D. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.