Local vs. Global Interpretability: A Computational Complexity Perspective

Shahaf Bassan [1] **Guy Amir** [1] **Guy Katz** [1]

Abstract

The local and global interpretability of various ML models has been studied extensively in recent years. However, despite significant progress in the field, many known results remain informal or lack sufficient mathematical rigor. We propose a framework for bridging this gap, by using computational complexity theory to assess local and global perspectives of interpreting ML models. We begin by proposing proofs for two novel insights that are essential for our analysis: (i) a duality between local and global forms of explanations; and (ii) the inherent uniqueness of certain global explanation forms. We then use these insights to evaluate the complexity of computing explanations, across three model types representing the extremes of the interpretability spectrum: (i) linear models; (ii) decision trees; and (iii) neural networks. Our findings offer insights into both the local and global interpretability of these models. For instance, under standard complexity assumptions such as P $\neq$ NP, we prove that selecting *global* sufficient subsets in linear models is computationally harder than selecting *local* subsets. Interestingly, with neural networks and decision trees, the opposite is true: it is harder to carry out this task locally than globally. We believe that our findings demonstrate how examining explainability through a computational complexity lens can help us develop a more rigorous grasp of the inherent interpretability of ML models.

1. Introduction

Interpretability is becoming an increasingly important aspect of ML models, as it plays a key role in ensuring their safety, transparency and fairness (Doshi-Velez & Kim, 2017). The ML community has been studying two notions of interpretability: *global interpretability*, aimed at understanding the overall decision logic of an ML model; and *local interpretability*, aimed at understanding specific decisions made by that model (Zhang et al., 2021; Du et al., 2019). The correlation between a model's local and global interpretability levels is not always entirely evident. For instance, (Molnar, 2020) argues that while the weights linked to a linear classifier can be used in interpreting its local decisions, this may not be the case for its global behavior.

Despite significant progress in ML interpretability techniques, there still remains a notable lack of mathematical rigor in our comprehension of the inherent interpretability of different ML models. The work of (Barceló et al., 2020) proposes addressing this gap by analyzing interpretability through the perspective of *computational complexity* theory. There, the goal is to deepen our understanding of interpretability by exploring the computational complexity involved in generating different kinds of explanations for various ML models. A model is considered interpretable if an explanation can be computed efficiently; and conversely, if deriving an explanation is computationally intractable, the model is regarded as uninterpretable.

The study on the complexity of obtaining explanations includes various ML models and diverse forms of explanations (Marques-Silva et al., 2020; Arenas et al., 2021a; 2022; Wäldchen et al., 2021; Marques-Silva et al., 2021). However, this prior work focused mainly on *local* forms of explanations — enabling the formal analysis of *local* interpretability across various contexts, rather than addressing the overarching global interpretability of these models.

Our contributions. We present a formal computational-complexity-based framework for evaluating both the local and global interpretability levels of ML models. We do this by analyzing the complexity associated with computing different forms of explanation, distinguishing between those that are local (specific to a particular instance **x**) and global (applicable to any potential instance **x**).

Our study focuses on the analysis of *formal* notions of explanations that satisfy logical and mathematical guarantees — a sub-field often referred to as *formal explainable*

[1]The Hebrew University of Jerusalem, Jerusalem, Israel. Correspondence to: Shahaf Bassan <shahaf.bassan@mail.huji.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

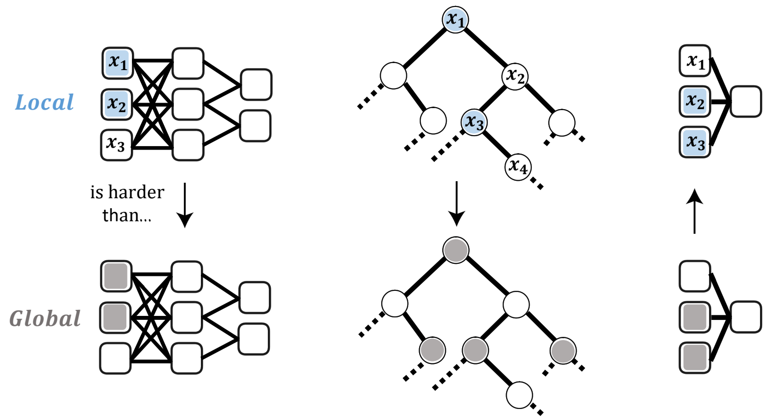

Figure 1. Illustration of complexity separations between local and global explanations. In linear models, it is harder to identify the smallest *global* sufficient subset (highlighted in gray) compared to a local one (highlighted in blue). Interestingly, this reverses in neural networks and decision trees, where selecting the smallest *global* sufficient subsets is computationally simpler than finding the smallest *local* ones.

AI (Marques-Silva & Ignatiev, 2022). The ability to deliver explanations with mathematically provable guarantees is crucial in safety-critical systems, and it also makes it possible to rigorously assess the computational complexity of obtaining such explanations. In this context, we focus on a few, commonly used formal notions of explanations:

1. **Sufficient Reason Feature Selection.** In the feature selection setting, users typically choose the k most significant input features. This selection can be executed either locally (selecting features that influence a particular prediction) or globally (selecting features that affect all instances within the domain). We examine the widely recognized *sufficiency* criterion, and explore the complexity of selecting subsets of features with the *smallest possible cardinality* while still maintaining sufficiency.

2. **Necessary and Redundant Features.** We analyze the computational complexity involved in identifying features that are either highly important or highly redundant. This type of analysis can be carried out in either a local or a global setting.

3. **Completion Count.** We consider a relaxed version of the former explainability forms, which computes the *relative portion* of assignments that maintain a prediction, given that we fix some subset of features. This form relates to the *probability* of obtaining a prediction, and can also be computed either locally or globally.

While the complexity of some of the *local* variants of these explanation forms has been studied previously, we focus here on their global variants. As part of our analysis, we present two novel theoretical insights: (i) a *duality* between local and global forms of explanations; and (ii) a result on the *uniqueness* of global sufficiency-based explanations, in stark contrast to the *exponential abundance* of their local counterparts. Using these insights, we are able to establish hitherto unknown complexity results on three model types that are frequently mentioned in the literature as being at the extremities of the interpretability spectrum: (i) decision trees; (ii) linear models; and (iii) neural networks.

In some cases, our complexity results rigorously justify prior claims. For example, we establish that linear models are indeed easier to interpret locally than globally under some contexts (Molnar, 2020) — selecting local sufficient subsets in these models can be performed in polynomial time, but selecting global sufficient subsets is coNP-Complete.

In other cases, however, our results actually defy intuition. For example, we discover that selecting global sufficient subsets is more tractable than local sufficient subsets, both for neural networks and decision trees (see illustration in figure 1). In the case of decision trees, for instance, the global form of this task can be performed in polynomial time, but the task becomes NP-Complete for its local counterpart. A similar phenomenon occurs with respect to identifying redundant features — this task is computationally harder to perform locally than globally, both for neural networks and decision trees.

We believe that these findings underscore the importance of rigorously analyzing the complexity of obtaining explanations, in order to enhance our understanding of model interpretability. While our study, like others in this field, is constrained by the specific explanation forms evaluated, we believe it provides a solid basis for deeper insights into both local and global interpretability, and paves the way for future exploration of additional explanation forms.

Due to space limitations, we provide a brief outline of the proofs for some of our claims within the paper, while the full proofs for all claims are relegated to the appendix.

2. Preliminaries

Complexity Classes. The paper assumes basic familiarity with the common complexity classes of polynomial time (PTIME) and nondeterministic polynomial time (NP, co-NP). We also mention classes of the second order of polynomial hierarchy, i.e., Σ_2^P, which describes the set of problems that could be solved in NP given an oracle that solves problems of co-NP in constant time, and Π_2^P, which describes the set of problems that could be solved in co-NP given an oracle that solves problems of NP in constant time. Both NP and co-NP are contained in both Σ_2^P and Π_2^P, and it is also widely believed that this containment is strict i.e., PTIME$\subsetneq$ NP, co-NP$\subsetneq$ Σ_2^P, Π_2^P (Arora & Barak, 2009). We also discuss the class #P, which corresponds to the total number of accepting paths of a polynomial-time nondeterministic Turing machine. It is also widely believed that #P strictly contains the second order of the polynomial hierarchy, i.e., that Σ_2^P, $\Pi_2^P \subsetneq$ #P (Arora & Barak, 2009).

Setting. We assume a set of n input feature assignments $\mathbf{x} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and use $\mathbb{F} := \{0,1\}^n$ to denote the entire feature space. Our goal is to interpret the prediction of a classifier $f : \mathbb{F} \to \{0,1\}$. In the *local* case, we seek the reason behind the prediction $f(\mathbf{x})$ for a specific instance $\mathbf{x}$. In the *global* case, we seek to explain the general behavior of f. We follow common practice and use boolean input and output values to simplify the presentation (Arenas et al., 2021a; Wäldchen et al., 2021; Barceló et al., 2020). We note, however, that many of our results carry over to the real-valued case as well; see appendix K for additional information.

Explainability Queries. To cope with the abstract nature of interpretability, prior work often uses a construct called an *explainability query* (Barceló et al., 2020), denoted Q, which defines an explanation of a specific type. As prior work focused mainly on *local* explanation forms, the input of Q is usually comprised of both f and a specific $\mathbf{x}$, and its output is an answer providing information regarding the interpretation of $f(\mathbf{x})$. For any given explainability query Q, we define its corresponding *global form of explanation* as G-Q. In contrast to Q, the input of G-Q does not include a specific instance $\mathbf{x}$, and the output conditions hold for *any* $\mathbf{x} \in \mathbb{F}$. We provide the full formalization of each local and global explainability query in Section 3.

3. Local and Global Explanation Forms

Although model interpretability is subjective, there are several commonly used notions of local and global explanations, on which we focus here:

Sufficient Reason Feature Selection. In the feature selection setting, it is common for users to choose the top k features participating in a model's decision. We consider the widely recognized *sufficeincy* criterion for this selection, which aligns with common explainability methods (Ribeiro et al., 2018; Carter et al., 2019; Ignatiev et al., 2019a). We follow common conventions and define a *local sufficient reason* as a subset of features, $S \subseteq \{1, \ldots, n\}$, such that when features in S are fixed to their corresponding values in $\mathbf{x}$, the prediction is determined to be $f(\mathbf{x})$, regardless of other features' assignments. Formally, S is a local sufficient reason with respect to $\langle f, \mathbf{x} \rangle$ iff it holds that:

$$\forall(\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})] \tag{1}$$

Here, $(\mathbf{x}_S; \mathbf{z}_{\bar{S}})$ represents an assignment where the values of elements of S are taken from $\mathbf{x}$, and the remaining values $\overline{S}$ are taken from $\mathbf{z}$.

In the *global feature selection* setting, it is common to choose the top features contributing to *all* instances (Wang et al., 2015). We define a set $S \subseteq \{1, \ldots, n\}$ as a *global sufficient reason* of f if it is a local sufficient reason for all $\mathbf{x}$. More formally:

$$\forall(\mathbf{x}, \mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})] \tag{2}$$

We denote $\text{suff}(f, \mathbf{x}, S) = 1$ when S is a local sufficient reason of $\langle f, \mathbf{x} \rangle$, and $\text{suff}(f, \mathbf{x}, S) = 0$ otherwise. Similarly, we denote $\text{suff}(f, S) = 1$ when S is a *global* sufficient reason of f, and $\text{suff}(f, S) = 0$ otherwise.

A common notion in the literature suggests that smaller sufficient reasons (i.e., with smaller $|S|$), whether local or global, are more meaningful than larger ones (Ribeiro et al., 2018; Carter et al., 2019; Ignatiev et al., 2019a; Halpern & Pearl, 2005). Consequently, it is common to consider the complexity of obtaining subsets of features of *minimal cardinality* (also known as *minimum sufficient reasons*). This leads us to our first explainability query:

> **MSR (Minimum Sufficient Reason)**:
> **Input**: Model f, input $\mathbf{x}$, and integer k
> **Output**: *Yes* if there exists some S such that $\text{suff}(f, \mathbf{x}, S) = 1$ and $|S| \leq k$, and *No* otherwise

To differentiate between the local and global setting, we use *G-MSR* to refer to the explainability query that obtains a cardinally minimal *global* sufficient reason of f. Due to space limitations, we relegate the full formalization of global queries to appendix A.

To better understand the complexity of the *MSR* and *G-MSR* queries, we also consider the analysis of a *refined* version of this query, which instead of obtaining a cardinally minimal sufficient reason, is concerned with simply checking whether a subset of features is a sufficient reason:

> **CSR (Check Sufficient Reason)**:
> **Input**: Model f, input $\mathbf{x}$, and subset of features S
> **Output**: *Yes* if suff$(f, \mathbf{x}, S) = 1$, and *No* otherwise

Similarly, *G-CSR* denotes the explainability query for checking whether a subset of features is a *global* sufficient reason. This formalization (along with all other global queries in this section) appears in appendix A.

Identifying Necessary and Redundant Features. When interpreting a model, it is common to measure the importance of each feature to a prediction. For a better understanding of the complexity of local and global computations, we consider here the complexity of identifying the two extreme cases: features that are either *necessary* or *redundant* to a prediction. We use the formal notation of necessity and redundancy proposed by (Huang et al., 2023); and note that this notation also aligns with other formal frameworks that deal with bias detection and fairness (Arenas et al., 2021a; Darwiche & Hirth, 2020; Ignatiev et al., 2020a). There, necessary features can be regarded as biased features, whereas redundant features are protected features, which should not be used for decision making — such as gender, age, etc. (see appendix L for more information).

Formally, we define feature i as *locally necessary* for $\langle f, \mathbf{x}\rangle$ if it is contained in *all* sufficient reasons of $\langle f, \mathbf{x}\rangle$. Equivalently, removing i from any sufficient reason S causes it to cease being sufficient; i.e., for any $S \subseteq \{1, \ldots, n\}$ it holds that suff$(f, \mathbf{x}, S) = 1 \rightarrow$ suff$(f, \mathbf{x}, S \setminus \{i\}) = 0$.

In the global case, we seek to determine whether i is *globally necessary* to f, meaning it is necessary to *all* instances of $\langle f, \mathbf{x}\rangle$. Formally, for any $\mathbf{x} \in \mathbb{F}$ and for any $S \subseteq \{1, \ldots, n\}$ it holds that suff$(f, \mathbf{x}, S) = 1 \rightarrow$ suff$(f, \mathbf{x}, S \setminus \{i\}) = 0$.

> **FN (Feature Necessity)**:
> **Input**: Model f, input $\mathbf{x}$, and integer i
> **Output**: *Yes* if i is necessary with respect to $\langle f, \mathbf{x}\rangle$, and *No* otherwise

Conversely, a feature i is termed *locally redundant* regarding $\langle f, \mathbf{x}\rangle$ if its removal from any sufficient reason S does not change S's sufficiency. Formally, for any $S \subseteq \{1, \ldots, n\}$ it holds that suff$(f, \mathbf{x}, S) = 1 \rightarrow$ suff$(f, \mathbf{x}, S \setminus \{i\}) = 1$. This is equivalent to i not being contained in any *minimal* sufficient reason.

> **FR (Feature Redundancy)**:
> **Input**: Model f, input $\mathbf{x}$, and integer i.
> **Output**: *Yes*, if i is redundant with respect to $\langle f, \mathbf{x}\rangle$, and *No* otherwise.

We say that a feature is *globally redundant* if it is locally redundant with respect to all inputs; i.e., for any $\mathbf{x} \in \mathbb{F}$ and $S \subseteq \{1, \ldots, n\}$ it holds that suff$(f, \mathbf{x}, S) = 1 \rightarrow$ suff$(f, \mathbf{x}, S \setminus \{i\}) = 1$.

Counting completions. Lastly, we explore a *relaxed* version of the previous forms, which is commonly analyzed in other formal frameworks (Barceló et al., 2020; Wäldchen et al., 2021; Izza et al., 2021). This explanation form is based on exploring the relative portion of assignment completions that maintain a specific classification. This relates to the *probability* that a prediction remains the same, assuming the other features are uniformly and independently distributed. We define the local completion count c of S as the relative portion of completions which maintain the prediction of $f(\mathbf{x})$:

$$c(S, f, \mathbf{x}) := \frac{|\{\mathbf{z} \in \{0,1\}^{|\overline{S}|}, f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})\}|}{|\{\mathbf{z} \in \{0,1\}^{|\overline{S}|}|} \quad (3)$$

In the global completion count case, we count the number of completions for all possible assignments $\mathbf{x} \in \mathbb{F}$:

$$c(S, f) := \frac{|\{\mathbf{x} \in \mathbb{F}, \mathbf{z} \in \{0,1\}^{|\overline{S}|}, f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})\}|}{|\{\mathbf{x} \in \mathbb{F}, \mathbf{z} \in \{0,1\}^{|\overline{S}|}|} \quad (4)$$

> **CC (Count Completions)**:
> **Input**: Model f, input x, and subset of features S
> **Output**: The completion count $c(S, f, \mathbf{x})$

We acknowledge that other explanation forms can be used, and do not argue that one form is superior to others; rather, our goal is to study some local and global versions of common explanation forms as a means of assessing the local and global interpretability of different ML models.

4. Properties of Global Explanations

We now present several novel results concerning the characteristics of the aforementioned local and global forms of explanation. Subsequently, in Section 5 we illustrate how these results significantly affect the complexity of computing such explanations.

4.1. Duality of Local and Global Explanations

Our analysis shows that there exists a dual relationship between local and global explanations. To better understand this relationship, we make use of the definition of *contrastive reasons*, which describes subsets of features that, when altered, may cause the classification to change. Formally, a subset of features S is a contrastive reason with respect to $\langle f, \mathbf{x}\rangle$ iff there exists some $\mathbf{z} \in \mathbb{F}$ such that $f(\mathbf{x}_{\bar{S}}; \mathbf{z}_S) \neq f(\mathbf{x})$.

While sufficient reasons provide answers to "*why?*" questions, i.e., "why was $f(\mathbf{x})$ classified to class i?", contrastive reasons seek to provide answers to "*why not?*" questions. Clearly, S is a sufficient reason of $\langle f, \mathbf{x}\rangle$ iff $\overline{S}$ is *not* a contrastive reason of $\langle f, \mathbf{x}\rangle$. Contrastive reasons are also well related to necessity. This is shown by the following theorem, whose proof appears in appendix C:

Theorem 1 *A feature i is necessary with respect to $\langle f, \boldsymbol{x}\rangle$ if and only if $\{i\}$ is a contrastive reason of $\langle f, \boldsymbol{x}\rangle$.*

We can similarly define a *global contrastive reason* as a subset of features that may cause a misclassification for any possible input. Formally, for any $\mathbf{x} \in \mathbb{F}$ there exists some $\mathbf{z} \in \mathbb{F}$ such that $f(\mathbf{x}_{\bar{S}}; \mathbf{z}_S) \neq f(\mathbf{x})$. This leads to a first dual relationship between local and global explanations:

Theorem 2 *Any global sufficient reason of f intersects with all local contrastive reasons of $\langle f, \boldsymbol{x}\rangle$, and any global contrastive reason of f intersects with all local sufficient reasons of $\langle f, \boldsymbol{x}\rangle$.*

This formulation can alternatively be expressed through the concept of *hitting sets* (additional details appear in appendix C). In this context, global sufficient reasons correspond to hitting sets of local contrastive reasons, while local contrastive reasons correspond to hitting sets for global sufficient reasons. It follows that the minimum hitting set (MHS; see appendix C) aligns with cardinally minimal reasons. Formally:

Theorem 3 *The MHS of all local contrastive reasons of $\langle f, \boldsymbol{x}\rangle$ is a cardinally minimal global sufficient reason of f, and the MHS of all local sufficient reasons of $\langle f, \boldsymbol{x}\rangle$ is a cardinally minimal global contrastive reason of f.*

For instance, suppose the set of all local contrastive reasons of f is $\mathbb{C} := \{\{1, 2\}, \{2, 3, 4\}, \{4, 5, 6\}\}$ (these may correspond to different local inputs $\mathbf{x} \in \mathbb{F}$). The smallest set intersecting all subsets in $\mathbb{C}$ (hence representing its MHS) is $\{2, 4\}$, thus $\{2, 4\}$ is the cardinally minimal *global* sufficient reason for f. Conversely, if $\mathbb{C}$ represents all local sufficient reasons, $\{2, 4\}$ represents the cardinally minimal global contrastive reason for f.

4.2. Uniquness of Global Explanations

As stated earlier, small sufficient reasons are often assumed to provide a better interpretation than larger ones. Consequently, we are interested in *minimal* sufficient reasons, i.e., explanation sets that cease to be sufficient reasons as soon as even one feature is removed from them. We note that *minimal* sufficient reasons are not necessarily cardinally minimal, and we can also consider *subset minimal* sufficient reasons (alternatively referred to as locally minimal). The choice of the terms *cardinally minimal* and *subset minimal* is deliberate, to reduce confusion with the concepts of global and local explanations.

A greedy approach for computing subset minimal sufficient reasons appears in Algorithm 1 (similar schemes appear in (Ignatiev et al., 2019a) and (Bassan & Katz, 2023)). It starts with the entire set of features, and then gradually attempts to remove features until converging to a *subset minimal* sufficient reason. Notably, the validation step at Line 3 within the algorithm, which determines the sufficiency of a feature subset, is not straightforward. In Section 5, we delve into a detailed discussion of the computational complexities associated with this process.

Algorithm 1 Local Subset Minimal Sufficient Reason

Input f, $\mathbf{x}$

1: $S \leftarrow \{1, \ldots, n\}$
2: **for each** $i \in \{1, ..., n\}$ by some arbitrary ordering **do**
3:     **if** suff$(f, \mathbf{x}, S \setminus \{i\}) = 1$ **then**
4:         $S \leftarrow S \setminus \{i\}$
5:     **end if**
6: **end for**
7: **return** S ▷ S is a subset minimal *local* sufficient reason

While Algorithm 1 converges to a subset-minimal local sufficient reason, it is not necessarily a cardinally minimal sufficient reason. This is due to the algorithm's strong sensitivity to the order in which we iterate over features (Line 2). The number of subset-minimal and cardinally minimal sufficient reasons depends on the function f. Nevertheless, it can be shown that their prevalence is, in the worst case, *exponential* in the number of features n:

Proposition 1 *There exists some f and some $\boldsymbol{x} \in \mathbb{F}$ such that there are $\Theta(\frac{2^n}{\sqrt{n}})$ local subset minimal or cardinally minimal sufficient reasons of $\langle f, \boldsymbol{x}\rangle$.*

A similar, greedy approach for computing subset minimal *global* sufficient reasons appears in Algorithm 2:

Algorithm 2 Global Subset Minimal Sufficient Reason

Input f

1: $S \leftarrow \{1, \ldots, n\}$
2: **for each** $i \in \{1, ..., n\}$ by some arbitrary ordering **do**
3:     **if** suff$(f, S \setminus \{i\}) = 1$ **then**
4:         $S \leftarrow S \setminus \{i\}$
5:     **end if**
6: **end for**
7: **return** S ▷ S is a subset minimal *global* sufficient reason

Given that the criteria for a subset of features to constitute a *global* sufficient reason are more stringent than those for the local case, it is natural to ask whether they are also exponentially abundant. To start addressing this question, we establish the following proposition:

Proposition 2 *If S_1 and S_2 are two global sufficient reasons of some non-trivial function f, then $S_1 \cap S_2 = S \neq \emptyset$, and S is a global sufficient reason of f.*

From Proposition 2 now stems the following theorem:

Theorem 4 *There exists one unique subset-minimal global sufficient reason of f.*

Thus, while the local form of explanation presents us with a worst-case scenario of an *exponential* number of minimal explanations, the global form, on the other hand, offers only a single, unique minimal explanation. As we demonstrate later, this distinction causes significant differences in the complexity of computing such explanations. We can now derive the following corollary:

Proposition 3 *For any possible ordering of features in Line 2 of Algorithm 2, Algorithm 2 converges to the same global sufficient reason.*

The uniqueness of global subset-minimal sufficient reasons also carries implications for the assessment of feature necessity and redundancy, as follows:

Proposition 4 *Let S be the subset minimal global sufficient reason of f. For all i, $i \in S$ if and only if i is locally necessary for some $\langle f, \boldsymbol{x} \rangle$, and $i \in \overline{S}$ if and only if i is globally redundant for f.*

In other words, subset S, which is the unique minimal global sufficient reason of f, categorizes the features into two possible sets: those *necessary* to a specific instance **x**, and those that are *globally redundant*. This fact is further exemplified by the subsequent corollary:

Proposition 5 *Any feature i is either locally necessary for some $\langle f, \boldsymbol{x} \rangle$, or globally redundant for f.*

The proofs for all propositions and theorems discussed in this section can be found in appendix C for claims related to duality, and in appendix D for those concerning uniqueness.

5. The Computational Complexity of Global Interpretation

We seek to comprehensively analyze the computational complexity of producing local and global explanations, of the forms discussed in Section 3. We perform this analysis on three classes of models: free binary decision diagrams (FBDDs), which are a generalization of decision trees; Perceptrons; and Multi-Layer Perceptrons (MLPs) with ReLU activation units. A full formalization of these model classes is provided in appendix B.

We use $Q(\mathcal{C})$ (respectively, G-$Q(\mathcal{C})$) to denote the computational problem of solving the *local* (respectively, *global*) explainability query Q on models of class $\mathcal{C}$. Table 1 summarizes our results, and indicates the complexity classes for model class and explanation type pairs.

As these results demonstrate, there is often a *strict disparity* in computational effort between calculating local and global explanations, emphasizing the need for distinct assessments of local and global forms. We further study this disparity and investigate the *comparative* computational efforts required for local and global explanations across various models and forms of explanations. This examination enables us to address the fundamental question of whether certain models exhibit a higher degree of interpretability at a global level compared to their interpretability at a local level, within different contextual scenarios.

Local vs. Global Interpretability

We say that a model is more *locally interpretable* for a given explanation type if computing the local form of that explanation is strictly easier than computing the global form, and say that it is more *globally interpretable* in the opposite case. We use the notation of *c-interpretablity* (computational interpretability) (Barceló et al., 2020) to study the disparity between local and global computations of our analyzed query forms. More formally:

Definition 1 *Let Q denote an explainability query and $\mathcal{C}$ a class of models, and suppose $Q(\mathcal{C})$ is in class $\mathcal{K}_1$ and G-$Q(\mathcal{C})$ is in class $\mathcal{K}_2$. Then:*

1. *$\mathcal{C}$ is strictly more locally c-interpretable with respect to Q iff $\mathcal{K}_1 \subsetneq \mathcal{K}_2$ and G-$Q(\mathcal{C})$ is hard for $\mathcal{K}_2$.*

2. *$\mathcal{C}$ is strictly more globally c-interpretable with respect to Q iff $\mathcal{K}_2 \subsetneq \mathcal{K}_1$ and $Q(\mathcal{C})$ is hard for $\mathcal{K}_1$.*

We divide our discussion into scenarios where computing an explanation is more challenging in the local setting, in the global setting, or equally difficult in both settings.

5.1. The Locally Interpretable Case

We start with the Perceptron model, where a strict complexity gap exists between local and global computations. Our findings reveal that in linear models, global feature selection is computationally harder than local feature selection. As shown in Table 1, there is a disparity in feature selection queries (*CSR* and *MSR*) between local and global forms. Local forms are achievable in polynomial time, whereas global forms are coNP-Complete, leading to our first corollary:

Table 1. Complexity classes for pairs of explainability queries and model classes. Cells highlighted in blue represent novel results, presented here; whereas the remaining results were already known previously.

	FBDDs		**MLPs**		**Perceptrons**	
	Local	Global	Local	Global	Local	Global
CSR	PTIME	PTIME	coNP-C	coNP-C	PTIME	coNP-C
MSR	NP-C	PTIME	Σ_2^P-C	coNP-C	PTIME	coNP-C
CC	PTIME	PTIME	#P-C	#P-C	#P-C	#P-C
FR	coNP-C	PTIME	Π_2^P-C	coNP-C	coNP-C	coNP-C
FN	PTIME	PTIME	PTIME	coNP-C	PTIME	PTIME

Theorem 5 *Perceptrons are strictly more locally c-interpretable with respect to CSR and MSR.*

The complexity difference arises from the intrinsic properties of linear models. Previous work showed that in linear models, selecting cardinally minimal *local* sufficient reasons can be performed in polynomial time (Barceló et al., 2020; Marques-Silva et al., 2020). The feasibility of such algorithms stems from the understanding that the local value $\mathbf{x}_i$ of a feature i, along with its corresponding weight $\mathbf{w}_i$, can be utilized to calculate the exact contribution of feature i to the local prediction $f(\mathbf{x})$.

However, we are able to prove that for *global* sufficient reasons, the situation is different, as the sufficiency criterion for the selection process considers all inputs $\mathbf{x}$, rather than a specific $\mathbf{x}$. This property makes the task of selecting global sufficient reasons for linear models intractable:

Proposition 6 *For Perceptrons, solving G-CSR and G-MSR is coNP-Complete, while solving CSR and MSR can be done in polynomial time.*

Proof Sketch. We prove in appendix E that for *G-CSR*, membership in coNP holds from guessing certificates $\mathbf{x} \in \mathbb{F}$ and $\mathbf{z} \in \mathbb{F}$ and validating whether S is not sufficient. For *G-MSR*, membership is a consequence of Proposition 4, which shows that any feature that is contained in the subset minimal global sufficient reason is necessary for some $\langle f, \mathbf{x} \rangle$, or is globally redundant otherwise. Hence, we can guess n assignments $\mathbf{x}^1, \ldots, \mathbf{x}^n$, and for each feature $i \in \{1, \ldots n\}$, validate whether i is locally necessary for $\langle f, \mathbf{x}^i \rangle$, and whether this holds for more than k features. We prove hardness with similar reductions for both *G-CSR* and *G-MSR* using a reduction from the $\overline{SSP}$ (subset-sum problem), which is coNP-Complete.

This result is particularly interesting since it provides evidence of a *lack of global interpretability* in linear models, in contrast to their inherent local interpretability, supporting intuition raised by previous works (Molnar, 2020).

Another case where local computations were found to be strictly less complex than global ones is identifying necessary features in neural networks:

Theorem 6 *MLPs are strictly more locally c-interpretable with respect to FN.*

This disparity was only found in MLPs, as demonstrated in the following proposition:

Proposition 7 *For FBDDs and Perceptrons, FN and G-FN can be solved in polynomial time. However, for MLPs, FN can be solved in polynomial time, while solving G-FN is coNP-Complete.*

Proof Sketch. We prove in appendix F that membership in coNP can be obtained using Theorem 1. We then prove hardness for MLPs by reducing from *TAUT*, a classic coNP-Complete problem that checks whether a boolean formula is a tautology. For Perceptrons and FBDDs we suggest polynomial algorithms whose correctness is derived from Theorem 1.

5.2. The Globally Interpretable Case

The fact that a model is more locally interpretable may seem intuitive. Nevertheless, it is rather surprising that, in certain instances, we can prove that the *global explanation forms are easier to compute than the local forms*. We demonstrate that this is sometimes the case both in neural networks and in decision trees. Specifically, in these models, local feature selection is computationally harder than global feature selection. This contrasts with linear models, where the local variant of feature selection was easier than the global:

Theorem 7 *FBDDs and MLPs are strictly more globally c-interpretable with respect to MSR.*

Unlike linear models, decision trees and neural networks do not possess the characteristics that enable a polynomial selection of cardinally minimal *local* sufficient reasons, which is NP-Complete for decision trees, and even less tractable (Σ_2^P-Complete) for neural networks (Barceló et al., 2020). Intuitively, NP-Hardness in decision trees stems from the need to explore an exponential number of possible subsets to find the smallest one. In neural networks, this complexity increases further because not only are there exponentially many subsets to consider, but verifying if one specific subset is a sufficient reason is already coNP-Complete, leading to the overall Σ_2^P complexity.

However, we have demonstrated that both models have strictly lower complexity when addressing *global* sufficient reasons, primarily due to the *uniqueness* property of subset-minimal global sufficient reasons established in Theorem 4. This contrasts with their exponential abundance in the local version (Proposition 1). We note that the complexity of checking whether a subset is sufficient is akin to local and global computations (*CSR* and *G-CSR*). However, when obtaining *cardinally minimal* sufficient reasons (*MSR*, *G-MSR*) there exists a strict disparity between local and global computations since the complexity is obviously tied to the number of possible subset candidates. This difference renders the global query version simpler to compute than the local, due to the underlying uniqueness property:

Proposition 8 *For FBDDs, G-CSR and G-MSR can be solved in polynomial time, while MSR is NP-Complete. Moreover, in MLPs, solving G-CSR and G-MSR is coNP-Complete, while solving MSR is Σ_2^P-Complete.*

Proof Sketch. We prove these results in appendix G. For FBDDs, we provide polynomial algorithms for the global queries, based on the observation that one can identify global sufficient reasons by iterating over pairs of leaf nodes, instead of iterating over single leaf nodes (which is how we identify local sufficient reasons). For MLPs, membership trivially holds, and the same hardness results used for Perceptrons hold here as well.

On the practical side, this finding demonstrates that obtaining cardinally minimal *global* sufficient reasons is feasible in some cases. For decision trees, accomplishing this task is feasible within polynomial time. When it comes to neural networks, the task can be executed with a linear number of calls to a coNP oracle, such as neural network verification tools (Wang et al., 2021b; Brix et al., 2023; Wu et al., 2024a).

This contrasts sharply with the *local* variant of this problem, which is Σ_2^P complete and thus necessitates an exponentially large number of verification queries, making this task infeasible, even with the use of neural network verifiers.

The phenomenon observed in the *MSR* query, where global explanations are easier to compute for decision trees and neural networks, is mirrored in the feature-redundancy (*FR*) query: it is computationally harder to identify locally redundant features than globally redundant ones in these models:

Theorem 8 *FBDDs and MLPs are strictly more globally c-interpretable with respect to FR.*

The complexity of identifying redundant features is closely linked to that of the *MSR* query. Recall that identifying a redundant feature is akin to validating whether a feature is not part of any subset minimal sufficient reason. While this is computationally challenging in the local scenario due to the exponential number of subset-minimal *local* sufficient reasons, it becomes more tractable in the *global* context.

Proposition 9 *For FBDDs, G-FR can be solved in polynomial time, while solving FR is coNP-Complete. Moreover, in MLPs, solving G-FR is coNP-Complete, while solving FR is Π_2^P-Complete.*

These results imply an intriguing observation regarding the complexity of identifying local and global necessary and redundant features in the specific case of MLPs:

Observation 1 *For MLPs, global necessity (G-FN) is strictly harder than local necessity (FN), whereas global redundancy (G-FR) is strictly less hard than local redundancy (FR).*

Another interesting insight from the previous theorems is the comparison between MLPs and Perceptrons. Since Perceptrons are a specific case of MLPs with one layer, analyzing the complexity difference between them can provide insights into the influence of hidden layers on model intricacy. Our findings indicate that while hidden layers influence the local queries we examined, they do not impact the global queries.

Observation 2 *Obtaining CSR, MSR, and FR is strictly harder for MLPs compared to MLPs with no hidden layers. However, this disparity does not exist for G-CSR, G-MSR, and G-FR.*

5.3. The Equally Difficult Case

Finally, we explore the cases in which local and global computations are complete for the same complexity class. First, we demonstrate that detecting redundant features at both local and global levels in Perceptrons is coNP-Complete:

Proposition 10 *For Perceptrons, solving FR and G-FR are both coNP-Complete.*

This stands in sharp contrast with decision trees and neural networks, where the global structure of the query is strictly simpler than its local counterpart. This difference is attributable, again, to the intrinsic properties of linear models, which lead to a diminished level of global interpretability, as discussed in subsection 5.1.

Next, we show that in *all* of our analyzed models, the complexity of the count-completion (*CC*) query remains the same across the local and global versions of computation:

Proposition 11 *For FBDDs, both CC and G-CC can be solved in polynomial time. Moreover, for Perceptrons and MLPs, solving both CC and G-CC are $\#P$-Complete.*

Proof Sketch. We prove these complexity results in appendix J. Membership in $\#P$ is straightfowrad. For the hardness in the case of Perceptrons/MLPs, we reduce from (local) *CC* of Perceptrons which is $\#P$-complete. For FBDDs, we propose a polynomial algorithm.

We recall that the *CC* query is a *relaxed* counting version of the *CSR* query, which seeks the relative portion of a subset S, instead of posing a decision problem about S. This complexity result highlights a notable distinction between decision problems (which exhibit a complexity gap between local and global forms) and counting problems, where such a complexity gap is absent.

6. Related Work

Our work contributes to *Formal XAI* (Marques-Silva et al., 2020), which focuses on the analysis of explanations with mathematical guarantees. Several papers have explored the computational complexity of obtaining such explanations (Barceló et al., 2020; Wäldchen et al., 2021; Arenas et al., 2022; 2021a;b; Ordyniak et al., 2023; Blanc et al., 2021; 2022; Van den Broeck et al., 2022; Marques-Silva et al., 2021; Audemard et al., 2020; Huang et al., 2022; Audemard et al., 2021; Izza & Marques-Silva, 2021); however, these notable efforts primarily focused on local forms of explanations, while our framework offers a comprehensive approach for analyzing and contrasting both local and global explanations.

Certain terms used in our work have been referred to by additional names in the literature: "sufficient reasons" are also known as *abductive explanations* (Ignatiev et al., 2019b), while minimal sufficient reasons are sometimes referred as *prime implicants* in Boolean formulas (Shih et al., 2018). A notion similar to the *CC* query is the δ-relevant set (Wäldchen et al., 2021; Izza et al., 2021), which asks whether the completion count exceeds a threshold δ. Similar duality properties to the ones studied here were shown to hold considering the relationship between local sufficient and contrastive reasons (Ignatiev et al., 2020b), and between absolute sufficient reasons and adversarial attacks (Ignatiev et al., 2019b). Minimal *absolute* sufficient reasons are the smallest-sized subsets among all possible inputs for a specific prediction and rely on partial input assignments. In our global sufficient reason definition, we do not rely on particular inputs or partial assignments (see appendix L for more details).

The necessity and redundancy queries that we discussed were studied previously (Huang et al., 2023) and were also explored under the context of bias detection (Arenas et al., 2021a; Darwiche & Hirth, 2020; Ignatiev et al., 2020a). We acknowledge, of course, that there exist many other notions of bias and fairness (Mehrabi et al., 2021).

7. Conclusion

We present a theoretical framework using computational complexity theory to assess both local and global perspectives of interpreting ML models. Our work uncovers new insights, including a duality between local and global explanations and the uniqueness inherent in some global explanation forms. We then build upon these insights and propose novel proofs for complexity classes tied to various explanation forms, enabling us to *formally measure interpretability* across different local and global contexts. While some of our findings justify folklore claims, others are unexpected. We believe that these discoveries illustrate the importance of applying computational complexity theory to gain a thorough understanding of the interpretability of ML models, paving the way for further research.

Impact Statement

While we acknowledge the potential social implications of interpretability, our work primarily focuses on theoretical aspects. Hence, we believe that it does not entail any direct social ramifications.

Acknowledgements

This work was partially funded by the European Union (ERC, VeriDeL, 101112713). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The work of Amir was further supported by a scholarship from the Clore Israel Foundation.

References

Amir, G., Schapira, M., and Katz, G. Towards Scalable Verification of Deep Reinforcement Learning. In *Proc. 21st Int. Conf. on Formal Methods in Computer-Aided Design (FMCAD)*, pp. 193–203, 2021a.

Amir, G., Wu, H., Barrett, C., and Katz, G. An SMT-Based Approach for Verifying Binarized Neural Networks. In

Proc. 27th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS), pp. 203–222, 2021b.

Amir, G., Zelazny, T., Katz, G., and Schapira, M. Verification-Aided Deep Ensemble Selection. In *Proc. 22nd Int. Conf. on Formal Methods in Computer-Aided Design (FMCAD)*, pp. 27–37, 2022.

Amir, G., Corsi, D., Yerushalmi, R., Marzari, L., Harel, D., Farinelli, A., and Katz, G. Verifying Learning-Based Robotic Navigation Systems. In *Proc. 29th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pp. 607–627, 2023a.

Amir, G., Freund, Z., Katz, G., Mandelbaum, E., and Refaeli, I. veriFIRE: Verifying an Industrial, Learning-Based Wildfire Detection System. In *Proc. 25th Int. Symposium on Formal Methods (FM)*, pp. 648–656, 2023b.

Amir, G., Maayan, O., Zelazny, T., Katz, G., and Schapira, M. Verifying Generalization in Deep Learning. In *Proc. of the 35th Int. Conf. on Computer Aided Verification (CAV)*, pp. 438–455, 2023c.

Arenas, M., Baez, D., Barceló, P., Pérez, J., and Subercaseaux, B. Foundations of Symbolic Languages for Model Interpretability. In *Proc. 34th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11690–11701, 2021a.

Arenas, M., Barceló, P., Bertossi, L., and Monet, M. The Tractability of SHAP-Score-Based Explanations for Classification Over Deterministic and Decomposable Boolean Circuits. In *Proc. 35th AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 6670–6678, 2021b.

Arenas, M., Barceló, P., Romero, M., and Subercaseaux, B. On Computing Probabilistic Explanations for Decision Trees. In *Proc. 35th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28695–28707, 2022.

Arora, S. and Barak, B. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

Audemard, G., Koriche, F., and Marquis, P. On Tractable XAI Queries Based on Compiled Representations. In *Proc. Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, pp. 838–849, 2020.

Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J., and Marquis, P. On the Computational Intelligibility of Boolean Classifiers. In *Proc. 18th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, pp. 74–86, 2021.

Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., and Marquis, P. On Preferred Abductive Explanations for Decision Trees and Random Forests. In *Proc. 31st Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 643–650, 2022.

Audemard, G., Lagniez, J.-M., Marquis, P., and Szczepanski, N. Computing Abductive Explanations for Boosted Trees. In *Proc. of the Int. Conf. on Artificial Intelligence and Statistics*, pp. 4699–4711, 2023.

Bacchus, F. and Katsirelos, G. Using Minimal Correction Sets to More Efficiently Compute Minimal Unsatisfiable Sets. In *Proc. Int. Conf. on Computer Aided Verification (CAV)*, pp. 70–86, 2015.

Bailey, J. and Stuckey, P. J. Discovery of Minimal Unsatisfiable Subsets of Constraints Using Hitting Set Dualization. In *Proc. 7th Int. Symposium on Practical Aspects of Declarative Languages (PADL)*, pp. 174–186, 2005.

Barceló, P., Monet, M., Pérez, J., and Subercaseaux, B. Model Interpretability through the Lens of Computational Complexity. In *Proc. 33rd Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15487–15498, 2020.

Bassan, S. and Katz, G. Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks. In *Proc. 29th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pp. 187–207, 2023.

Bassan, S., Amir, G., Corsi, D., Refaeli, I., and Katz, G. Formally Explaining Neural Networks within Reactive Systems. In *Proc. 23rd Int. Conf. on Formal Methods in Computer-Aided Design (FMCAD)*, pp. 10–22, 2023.

Birnbaum, E. and Lozinskii, E. Consistent Subsets of Inconsistent Systems: Structure and Behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, 15(1): 25–46, 2003.

Blanc, G., Lange, J., and Tan, L.-Y. Provably Efficient, Succinct, and Precise Explanations. In *Proc. 35th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6129–6141, 2021.

Blanc, G., Koch, C., Lange, J., and Tan, L.-Y. A Query-Optimal Algorithm for Finding Counterfactuals. In *Proc. 39th Int. Conf. on Machine Learning (ICML)*, pp. 2075–2090, 2022.

Boumazouza, R., Cheikh-Alili, F., Mazure, B., and Tabia, K. ASTERYX: A model-Agnostic SaT-basEd appRoach for SYmbolic and Score-Based EXplanations. In *Proc. 30th ACM Int. Conf. on Information & Knowledge Management (CIKM)*, pp. 120–129, 2021.

Brix, C., Müller, M., Bak, S., Johnson, T., and Liu, C. First Three Years of the International Verification of Neural Networks Competition (VNN-COMP). *International Journal on Software Tools for Technology Transfer*, pp. 1–11, 2023.

Carter, B., Mueller, J., Jain, S., and Gifford, D. What Made You Do This? Understanding Black-Box Decisions with Sufficient Input Subsets. In *Proc. 22nd Int. Conf. on Artificial Intelligence and Statistics (AIStat)*, pp. 567–576, 2019.

Casadio, M., Komendantskaya, E., Daggitt, M., Kokke, W., Katz, G., Amir, G., and Refaeli, I. Neural Network Robustness as a Verification Property: A Principled Case Study. In *Proc. 34th Int. Conf. on Computer Aided Verification (CAV)*, pp. 219–231, 2022.

Corsi, D., Yerushalmi, R., Amir, G., Farinelli, A., Harel, D., and Katz, G. Constrained Reinforcement Learning for Robotics via Scenario-Based Programming, 2022. Technical Report. https://arxiv.org/abs/2206.09603.

Corsi, D., Amir, G., Katz, G., and Farinelli, A. Analyzing Adversarial Inputs in Deep Reinforcement Learning, 2024a. Technical Report. https://arxiv.org/abs/2402.05284.

Corsi, D., Amir, G., Rodríguez, A., Sánchez, C., Katz, G., and Fox, R. Verification-Guided Shielding for Deep Reinforcement Learning. In *Proc. 1st Int. Reinforcement Learning Conf. (RLC)*, 2024b.

Darwiche, A. Logic for Explainable AI. In *Proc. 38th ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pp. 1–11, 2023.

Darwiche, A. and Hirth, A. On the Reasons Behind Decisions. In *Proc. 24th European Conf. on Artificial Intelligence (ECAI)*, pp. 712–720, 2020.

Darwiche, A. and Ji, C. On the Computation of Necessary and Sufficient Explanations. In *Proc. 36th AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 5582–5591, 2022.

De Moura, L. and Bjørner, N. Z3: An Efficient SMT Solver. In *Proc. Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pp. 337–340, 2008.

Doshi-Velez, F. and Kim, B. Towards a Rigorous Science of Interpretable Machine Learning, 2017. Technical Report. https://arxiv.org/abs/1702.08608.

Du, M., Liu, N., and Hu, X. Techniques for Interpretable Machine Learning. *Communications of the ACM*, 63(1): 68–77, 2019.

Fel, T., Ducoffe, M., Vigouroux, D., Cadène, R., Capelle, M., Nicodème, C., and Serre, T. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 16153–16163, 2023.

Gardner, M. and Dorling, S. Artificial Neural Networks (the Multilayer Perceptron) — A Review of Applications in the Atmospheric Sciences. *Atmospheric Environment*, 32 (14-15):2627–2636, 1998.

Gorji, N. and Rubin, S. Sufficient Reasons for Classifier Decisions in the Presence of Domain Constraints. In *Proc. 36th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5660–5667, 2022.

Halpern, J. and Pearl, J. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 2005.

Huang, X. and Marques-Silva, J. From Robustness to Explainability and Back Again, 2023. Technical Report. https://arxiv.org/abs/2306.03048.

Huang, X., Izza, Y., Ignatiev, A., and Marques-Silva, J. On Efficiently Explaining Graph-Based Classifiers. In *Proc. 18th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, 2021.

Huang, X., Izza, Y., Ignatiev, A., Cooper, M., Asher, N., and Marques-Silva, J. Tractable Explanations for d-DNNF Classifiers. In *Proc. 36th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5719–5728, 2022.

Huang, X., Cooper, M., Morgado, A., Planes, J., and Marques-Silva, J. Feature Necessity & Relevancy in ML Classifier Explanations. In *Proc. 29th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pp. 167–186, 2023.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. Abduction-Based Explanations for Machine Learning Models. In *Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1511–1519, 2019a.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. On Relating Explanations and Adversarial Examples. In *Proc. 32nd Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

Ignatiev, A., Cooper, M., Siala, M., Hebrard, E., and Marques-Silva, J. Towards Formal Fairness in Machine Learning. In *Proc. 26th Int. Conf. on Principles and Practice of Constraint Programming (CP)*, pp. 846–867, 2020a.

Ignatiev, A., Narodytska, N., Asher, N., and Marques-Silva, J. From Contrastive to Abductive Explanations and Back Again. In *Proc. Int. Conf. of the Italian Association for Artificial Intelligence (AIxAI)*, pp. 335–355, 2020b.

Ignatiev, A., Izza, Y., Stuckey, P. J., and Marques-Silva, J. Using MaxSAT for Efficient Explanations of Tree Ensembles. In *Proc. 36th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3776–3785, 2022.

Izza, Y. and Marques-Silva, J. On Explaining Random Forests with SAT. In *Proc. 30th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2021.

Izza, Y., Ignatiev, A., Narodytska, N., Cooper, M., and Marques-Silva, J. Efficient Explanations with Relevant Sets, 2021. Technical Report. https://arxiv.org/abs/2106.00546.

Izza, Y., Ignatiev, A., and Marques-Silva, J. On Tackling Explanation Redundancy in Decision Trees. *Journal of Artificial Intelligence Research*, 75:261–321, 2022.

Izza, Y., Huang, X., Morgado, A., Planes, J., Ignatiev, A., and Marques-Silva, J. Distance-Restricted Explanations: Theoretical Underpinnings & Efficient Implementation, 2024. Technical Report. https://arxiv.org/abs/2405.08297.

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *Proc. 29th Int. Conf. on Computer Aided Verification (CAV)*, pp. 97–117, 2017.

Lee, C. Representation of Switching Circuits by Binary-Decision Programs. *The Bell System Technical Journal*, 38(4):985–999, 1959.

Liffiton, M. H. and Sakallah, K. A. Algorithms for Computing Minimal Unsatisfiable Subsets of Constraints. *Journal of Automated Reasoning*, 40:1–33, 2008.

Liffiton, M. H., Previti, A., Malik, A., and Marques-Silva, J. Fast, Flexible MUS Enumeration. *Constraints*, 21: 223–250, 2016.

Malfa, E. L., Zbrzezny, A., Michelmore, R., Paoletti, N., and Kwiatkowska, M. On Guaranteed Optimal Robust Explanations for NLP Models. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2021.

Mandal, U., Amir, G., Wu, H., Daukantas, I., Newell, F., Ravaioli, U., Meng, B., Durling, M., Ganai, M., Shim, T., Katz, G., and Barrett, C. Formally Verifying Deep Reinforcement Learning Controllers with Lyapunov Barrier Certificates, 2024. Technical Report. https://arxiv.org/abs/2405.14058.

Marques-Silva, J. Logic-Based Explainability in Machine Learning. In *Proc. 18th Int. Summer School on Reasoning Web. Causality, Explanations and Declarative Knowledge*, pp. 24–104, 2023.

Marques-Silva, J. and Ignatiev, A. Delivering Trustworthy AI through formal XAI. In *Proc. 36th AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 12342–12350, 2022.

Marques-Silva, J., Gerspacher, T., Cooper, M., Ignatiev, A., and Narodytska, N. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *Proc. 33rd Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 20590–20600, 2020.

Marques-Silva, J., Gerspacher, T., Cooper, M., Ignatiev, A., and Narodytska, N. Explanations for Monotonic Classifiers. In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 7469–7479, 2021.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Molnar, C. *Interpretable Machine Learning*. Lulu.com, 2020.

Ordyniak, S., Paesani, G., and Szeider, S. The Parameterized Complexity of Finding Concise Local Explanations. In *Proc. 32nd Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2023.

Ralston, A., Reilly, E., and Hemmendinger, D. *Encyclopedia of Computer Science*. John Wiley and Sons Ltd., GBR, 2003. ISBN 0470864125.

Ramchoun, H., Ghanou, Y., Ettaouil, M., and Amine Janati Idrissi, M. Multilayer Perceptron: Architecture Optimization and Training. *Int. Journal of Interactive Multimedia and Artificial Intelligence*, 2016.

Ribeiro, M., Singh, S., and Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In *Proc. 32nd AAAI Conf. on Artificial Intelligence (AAAI)*, 2018.

Sälzer, M. and Lange, M. Reachability is NP-Complete Even for the Simplest Neural Networks. In *Proc. 15th Int. Conf. Reachability Problems (RP)*, pp. 149–164, 2021.

Shih, A., Choi, A., and Darwiche, A. A Symbolic Approach to Explaining Bayesian Network Classifiers. In *Proc. 27th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 5103–5111, 2018.

Van den Broeck, G., Lykov, A., Schleich, M., and Suciu, D. On the Tractability of SHAP Explanations. *Journal of Artificial Intelligence Research*, 74:851–886, 2022.

Wäldchen, S., Macdonald, J., Hauch, S., and Kutyniok, G. The Computational Complexity of Understanding Binary Classifier Decisions. *Journal of Artificial Intelligence Research (JAIR)*, 70:351–387, 2021.

Wang, D., Nie, F., and Huang, H. Feature Selection via Global Redundancy Minimization. *IEEE Transactions on Knowledge and Data Engineering*, 27(10):2743–2755, 2015.

Wang, E., Khosravi, P., and Van den Broeck, G. Probabilistic Sufficient Explanations. In *Proc. 30th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2021a.

Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C., and Kolter, Z. Beta-Crown: Efficient Bound Propagation with Per-Neuron Split Constraints for Neural Network Robustness Verification. In *Proc. 35th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 29909–29921, 2021b.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 1382–1392, 2021.

Wu, H., Isac, O., Zeljić, A., Tagomori, T., Daggitt, M., Kokke, W., Refaeli, I., Amir, G., Julian, K., Bassan, S., et al. Marabou 2.0: A Versatile Formal Analyzer of Neural Networks, 2024a. Technical Report. https://arxiv.org/abs/2401.14461.

Wu, M., Wu, H., and Barrett, C. Verix: Towards Verified Explainability of Deep Neural Networks. In *Proc. 36th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Yerushalmi, R., Amir, G., Elyasaf, A., Harel, D., Katz, G., and Marron, A. Scenario-Assisted Deep Reinforcement Learning. In *Proc. 10th Int. Conf. on Model-Driven Engineering and Software Development (MODELSWARD)*, pp. 310–319, 2022.

Yerushalmi, R., Amir, G., Elyasaf, A., Harel, D., Katz, G., and Marron, A. Enhancing Deep Reinforcement Learning with Scenario-Based Modeling. *SN Computer Science*, 4 (2):156, 2023.

Yu, J., Ignatiev, A., Stuckey, P., Narodytska, N., and Marques-Silva, J. Eliminating the Impossible, Whatever Remains Must Be True: On Extracting and Applying Background Knowledge in the Context of Formal Explanations. In *Proc. 37th AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 4123–4131, 2023.

Zhang, Y., Tiňo, P., Leonardis, A., and Tang, K. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5): 726–742, 2021.

Appendix

The appendix contains formalizations and proofs that were mentioned throughout the paper:

Appendix A formalizes the set of *global* explainability queries.
Appendix B formalizes the classes of models that were assessed in the paper.
Appendix C contains the proofs regarding the duality between local and global forms of explanations.
Appendix D contains the proofs concerning the inherent uniqueness of global forms of explanations.
Appendix E contains the proof of Proposition 6 (Complexity of *G-CSR* and *G-MSR* for Perceptrons)
Appendix F contains the proof of Proposition 7 (Complexity of *FN* and *G-FN* for FBDDs, Perceptrons, and MLPs)
Appendix G contains the proof of Proposition 8 (Complexity of *G-CSR* and *G-MSR* for FBDDs and MLPs)
Appendix H contains the proof of Proposition 9 (Complexity of *G-FR* for FBDDs and MLPs)
Appendix I contains the proof of Proposition 10 (Complexity of *FR* and *G-FR* for Perceptrons)
Appendix J contains the proof of Proposition 11 (Complexity of *G-CC* for FBDDs, Perceptrons, and MLPs)
Appendix K contains details on various extensions of this work.
Appendix L provides details on the terminologies and forms of explanation that are relevant to those discussed here.

A. Global forms of model explanations

In this section, we present the *global* forms of the explainability queries previously mentioned, which were initially formulated in the paper for their local configuration.

G-CSR (*Global* Check Sufficient Reason):
Input: Model f,and subset of features S.
Output: *Yes*, if S is a global sufficient reason of f (i.e., suff$(f, S) = 1$), and *No* otherwise.

G-MSR (*Global* Minimum Sufficient Reason):
Input: Model f, and integer k.
Output: *Yes*, if there exists a global sufficient reason S for f (i.e., suff$(f, S) = 1$) such that $|S| \leq k$, and *No* otherwise.

G-FR (*Global* Feature Redundancy):
Input: Model f, and integer i.
Output: *Yes*, if i is globally redundant with respect to f, and *No* otherwise.

G-FN (*Global* Feature Necessity):
Input: Model f, and integer i.
Output: *Yes*, if i is globally necessary with respect to f, and *No* otherwise.

G-CC (*Global* Count Completions):
Input: Model f, and subset S.
Output: The global completion count $c(S, f)$

B. Model Classes

Next, we describe in detail the various model classes that were taken into account within this work:

Free Binary Decision Diagram (FBDD). A *BDD* is a graph-based model that represents a Boolean function $f : \mathbb{F} \to \{0, 1\}$ (Lee, 1959). The arbitrary Boolean function is realized by an acyclic (directed) graph, for which the following holds: (i) every internal node v corresponds with a single binary input feature $(1, \ldots, n)$; (ii) every internal node v has exactly two output edges, that represent the values $\{0, 1\}$ assigned to v; (iii) each leaf node corresponds to either a *True*, or *False*, label; and (iv) every variable appears *at most* once, along every path α of the BDD.

Hence, any assignment to the inputs $\mathbf{x} \in \mathbb{F}$ corresponds to one unique path α from the BDD's root to one of its leaf nodes, and any path α corresponds to some partial assignment $\mathbf{x}_S$. We denote $f(\mathbf{x}) := 1$ if the label of the leaf node is true, and $f(\mathbf{x}) := 0$ if it is false. Moreover, we regard the size of a BDD (i.e., $|f|$) to be the overall number of edges in the BDD's graph. In this work, we focus on the popular variant of *Free BDD* (FBDD) models, for which different paths, α, α' are allowed to have different orderings of the input variables $\{1, \ldots, n\}$ and on every path α no two nodes have the same label. A *decision tree* can be essentially described as an FBDD whose foundational graph structure is a tree.

In our representation of the path α, we denote the nodes participating in the path as $\{\alpha_1, \alpha_2, \ldots, \alpha_t\}$, where t represents the total number of nodes in α. Each node α_i within the path α, is defined similarly to any general node in the FBDD. However, while any node possesses two output edges, we specify that a *path* node possesses a *single path* output edge. This edge corresponds to the value assigned to the binary feature associated with node α_i (either 1 or 0) in its corresponding partial assignment.

Multi-Layer Perceptron (MLP). A Multi-Layer Perceptron (Gardner & Dorling, 1998; Ramchoun et al., 2016) f with $t-1$ hidden layers (g^j for $j \in \{1, \ldots, t-1\}$) and a single output layer (g^t), is recursively defined as follows: $g^{(j)} := \sigma^{(j)}(g^{(j-1)}W^{(j)} + b^{(j)}) \quad (j \in \{1, \ldots, t\})$, given t weight matrices $W^{(1)}, \ldots, W^{(t)}$, t bias vectors $b^{(1)}, \ldots, b^{(t)}$, and also t activation functions $\sigma^{(1)}, \ldots, \sigma^{(t)}$.

The MLP f outputs the value $f := g^{(t)}$, while $g^{(0)} := \mathbf{x} \in \{0,1\}^n$ is the input layer that receives the input of the model. The biases and weight matrices are defined by a series of positive values $d_0, \ldots, d_t$ that represent their dimensions. Furthermore, we assume that all the weights and biases possess rational values, denoted as $W^{(j)} \in \mathbb{Q}^{d_{j-1} \times d_j}$ and $b^{(j)} \in \mathbb{Q}^{d_j}$, which have been acquired during the training phase. Due to our focus on *binary* classifiers over $\{1, \ldots, n\}$, it necessarily holds that: $d_0 = n$ and $d_t = 1$. In this work, we focus on the popular $ReLU(x) = \max(0, x)$ activation function, with the exception of the single activation in the last layer, that is typically a sigmoid function. Nonetheless, given our emphasis on post-hoc interpretations, it is without loss of generality that we may assume the last activation function is represented by the step function, i.e., $step(\mathbf{z}) = 1 \iff \mathbf{z} > 0$.

Perceptron. A Perceptron (Ralston et al., 2003) is a single-layered MLP (i.e., $t = 1$): $f(\mathbf{x}) = step((\mathbf{w} \cdot \mathbf{x}) + b)$, for $b \in \mathbb{Q}$ and $\mathbf{w} \in \mathbb{Q}^{n \times d_1}$. Thus, for a Perceptron f the following holds w.l.o.g.: $f(\mathbf{x}) = 1 \iff (\mathbf{w} \cdot \mathbf{x}) + b \geq 0$.

C. The Duality of Local and Global Explanations

Minimum Hitting Set (MHS). Given a collection $\mathbb{S}$ of sets from a universe U, a hitting set h for $\mathbb{S}$ is a set such that $\forall S \in \mathbb{S}, h \cap S \neq \emptyset$. A hitting set h is said to be *minimal* if none of its subsets is a hitting set, and *minimum* when it has the smallest possible cardinality among all hitting sets.

Theorem 1 *A feature i is necessary with respect to $\langle f, \mathbf{x} \rangle$ if and only if $\{i\}$ is a contrastive reason of $\langle f, \mathbf{x} \rangle$.*

Proof. For the first direction, let us begin by assuming that $\{i\}$ is a contrastive reason with respect to $\langle f, \mathbf{x} \rangle$. It then follows, from the definition of sufficient and contrastive reasons, that $\{1, \ldots, n\} \setminus \{i\}$ is *not* a sufficient reason for $\langle f, \mathbf{x} \rangle$. Consequently, any subset $S \subseteq \{1, \ldots, n\} \setminus \{i\}$ is also not a sufficient reason for $\langle f, \mathbf{x} \rangle$, which is equivalent to saying that for any $S \subseteq \{1, \ldots, n\}$ it holds that suff$(f, \mathbf{x}, S \setminus \{i\}) = 0$. This is true whether S is or is not a sufficient reason (and hence is particularly true for the case where it is one, i.e., suff$(f, \mathbf{x}, S) = 1$). As a direct consequence, for any $S \subseteq \{1, \ldots, n\}$ the following condition holds: suff$(f, \mathbf{x}, S) = 1 \rightarrow$ suff$(f, \mathbf{x}, S \setminus \{i\}) = 0$.

For the second direction, let us assume that i is necessary with respect to $\langle f, \mathbf{x} \rangle$, which particularly means that for all $S \subseteq \{1, \ldots, n\}$ it holds that suff$(f, \mathbf{x}, S) = 1 \rightarrow$ suff$(f, \mathbf{x}, S \setminus \{i\}) = 0$. We now assume, by contradiction, that $\{i\}$ is not a contrastive reason for $\langle f, \mathbf{x} \rangle$. Therefore, it follows, from the very definition of sufficient and contrastive reasons, that $\{1, \ldots, n\} \setminus \{i\}$ *is* a sufficient reason for $\langle f, \mathbf{x} \rangle$. Moreover, it clearly holds that the entire set $\{1, \ldots, n\}$ is a sufficient reason with respect to $\langle f, \mathbf{x} \rangle$, which is a property that holds for any f and any $\mathbf{x}$ (fixing *all* features necessarily determines that the prediction remains the same). Overall, we get that:

$$\text{suff}(f, \mathbf{x}, \{1, \ldots, n\}) = 1 \land \text{suff}(f, \mathbf{x}, \{1, \ldots, n\} \setminus \{i\}) = 1 \tag{5}$$

This is in contradiction to the assumption that i is necessary with respect to $\langle f, \mathbf{x} \rangle$.

Theorem 2 *Any global sufficient reason of f intersects with all local contrastive reasons of $\langle f, \mathbf{x} \rangle$ and any global contrastive reason of f intersects with all local sufficient reasons of $\langle f, \mathbf{x} \rangle$.*

Proof. For the first part, given some f and some $\mathbf{x}$, let us assume, by contradiction, that there exists some global sufficient reason S of f and some local contrastive reason S' of $\langle f, \mathbf{x}\rangle$ for which it holds that $S \cap S' = \emptyset$. Given that $S \cap S' = \emptyset$, it naturally follows that $S' \subseteq \overline{S}$. Taking into account that S is a global sufficient reason of f, we can infer that S is also a *local* sufficient reason of $\langle f, \mathbf{x}\rangle$. Therefore, from the definition of sufficient and contrastive reasons, $\overline{S}$ does not qualify as a contrastive reason with respect to $\langle f, \mathbf{x}\rangle$, leading to the implication that no subset of $\overline{S}$ can be a contrastive reason either. This assertion, however, contradicts the previously established $S' \subseteq \overline{S}$.

The second part of the claim will be almost identical to the first part: given some f and $\mathbf{x}$, we can again assume, by contradiction, that there exists some global *contrastive* reason S of f and some local *sufficient* reason S' of $\langle f, \mathbf{x}\rangle$ for which it holds that: $S \cap S' = \emptyset$. Given that $S \cap S' = \emptyset$, it naturally follows that $S' \subseteq \overline{S}$. Since S is a global contrastive reason of f it also acts as a local contrastive reason for $\langle f, \mathbf{x}\rangle$. As a consequence, from the very definition of sufficient and contrastive reasons, $\overline{S}$ can not be a sufficient reason for $\langle f, \mathbf{x}\rangle$. This implies that no subset of $\overline{S}$ can serve as a sufficient reason for $\langle f, \mathbf{x}\rangle$, creating a contradiction with the premise that $S' \subseteq \overline{S}$.

Theorem 3 *The MHS of all local contrastive reasons of $\langle f, \boldsymbol{x}\rangle$ is a cardinally minimal global sufficient reason of f, and the MHS of all local sufficient reasons of $\langle f, \boldsymbol{x}\rangle$ is a cardinally minimal global contrastive reason of f.*

Given some f, we denote $\mathbb{S}$ as the set of all local sufficient reasons of $\langle f, \mathbf{x}\rangle$ and denote $\mathbb{C}$ as the set of all local contrastive reasons of $\langle f, \mathbf{x}\rangle$. As a direct consequence of Theorem 2, we can determine the following claim:

Lemma 1 *A subset S is a global sufficient reason of f if and only if S is a hitting set of $\mathbb{S}$ and is a global contrastive reason of f if and only if S is a hitting set of $\mathbb{C}$.*

As a consequence of Lemma 1, it directly follows that cardinally minimal local contrastive reasons are with correspondence to MHSs of $\mathbb{S}$, and cardinally minimal local sufficient reasons are with correspondence to MHSs of $\mathbb{C}$.

The importance of the MHS duality. An essential finding when dealing with inconsistent sets of clauses lies in a similar MHS duality between Minimal Unsatisfiable Sets (MUSes) and Minimal Correction Sets (MCSes) (Birnbaum & Lozinskii, 2003; Bacchus & Katsirelos, 2015). In this context, MCSes are MHSs of MUSes, and vice versa (Bailey & Stuckey, 2005; Liffiton & Sakallah, 2008). This discovery has played a pivotal role in the advancement of algorithms designed for MUSes and MCSes and this result has found applications in various contexts (Bacchus & Katsirelos, 2015; Liffiton et al., 2016). While the majority of this research focuses on propositional theories, others focus on Satisfiability Modulo Theories (SMT) (De Moura & Bjørner, 2008).

Within the context of explainable AI, previous research has shown similar duality principles considering the relationship between *local* contrastive and sufficient reasons (Ignatiev et al., 2020b) as well as the relationship between absolute sufficient reasons and adversarial attacks (Ignatiev et al., 2019b). This relationship was shown to be critical in the exact computation of local sufficient reasons for various ML models such as decision trees (Izza et al., 2022), tree ensembles (Audemard et al., 2023; 2022; Izza & Marques-Silva, 2021; Ignatiev et al., 2022; Boumazouza et al., 2021), and neural networks (Bassan & Katz, 2023; Malfa et al., 2021). In neural networks, obtaining exact formal explanations presents substantial computational challenges. However, these explanations can be approximated with neural network verifiers, which are increasingly utilized for this purpose (Bassan et al., 2023; Bassan & Katz, 2023; Malfa et al., 2021; Wu et al., 2024b; Huang & Marques-Silva, 2023; Fel et al., 2023) and for formally verifying other properties (Casadio et al., 2022; Amir et al., 2021b; 2022; 2023b;c). More recently, (Bassan et al., 2023) demonstrated that leveraging the MHS duality may have an even larger effect of efficiency enhancement when computing sufficient reasons in neural networks, particularly within the domain of *reactive* systems, a domain in which formal analysis is extensively developed (Amir et al., 2021a; Corsi et al., 2022; Yerushalmi et al., 2022; 2023; Amir et al., 2023a; Corsi et al., 2024a;b; Mandal et al., 2024).

D. The Uniqueness of Global Explanations

Proposition 1 *There exists some f and some $\boldsymbol{x} \in \mathbb{F}$ such that there are $\Theta(\frac{2^n}{\sqrt{n}})$ local subset minimal or cardinally minimal sufficient reasons of $\langle f, \boldsymbol{x}\rangle$.*

Proof. We construct f as follows:

$$f(y) = \begin{cases} 1 & if \ \sum_{i=1}^{n} y_i \geq \lfloor \frac{n}{2} \rfloor \\ 0 & otherwise \end{cases} \tag{6}$$

We define the instance $\mathbf{x} := \mathbf{1}$. Clearly any subset S of size $\lfloor \frac{n}{2} \rfloor$ or larger is a local sufficient reason of $\langle f, \mathbf{x} \rangle$ (since fixing the values of S to $\mathbf{x}$ determines that the prediction remains: 1). Furthermore, every one of these subsets is *minimal* due to the fact that any subset of size $\lfloor \frac{n}{2} \rfloor - 1$ or smaller is *not* a sufficient reason of $\langle f, \mathbf{x} \rangle$ (it may cause a misclassification to class 0). Thus, it satisfies that there are $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ subset minimal local sufficient reasons of $\langle f, \mathbf{x} \rangle$. From Stirling's approximation, it holds that:

$$\lim_{n \to \infty} \frac{2\sqrt{2\pi}}{e^2} \cdot \frac{2^n}{\sqrt{n}} \leq \binom{n}{\lfloor \frac{n}{2} \rfloor} \leq \frac{e}{\pi} \cdot \frac{2^n}{\sqrt{n}} \tag{7}$$

This implies that $\binom{n}{\lfloor \frac{n}{2} \rfloor} = \Theta(\frac{2^n}{\sqrt{n}})$. Given that no local sufficient reason of size smaller than $\lfloor \frac{n}{2} \rfloor$ is present, these are, also *cardinally*-minimal sufficient reasons.

Proposition 2 *If S_1 and S_2 are two global sufficient reasons of some non-trivial f, then $S_1 \cap S_2 = S \neq \emptyset$, and S is a global sufficient reason of f.*

Proof. Our proof focuses on *non-trivial* functions, i.e., functions that do not always output 0 or always output 1. In other words, there exist some $\mathbf{x}, \mathbf{y} \in \mathbb{F}$ such that $f(\mathbf{x}) = 1$ and $f(\mathbf{y}) = 0$. However, it is important to point out that the general uniqueness property we will prove later (Theorem 4) is independent of whether f is non-trivial.

We begin by proving the following lemma:

Lemma 2 *For any f and $\mathbf{x} \in \mathbb{F}$, if S is a sufficient reason of $\langle f, \mathbf{x} \rangle$ then there does not exist any $\mathbf{y} \in \mathbb{F}$ such that $f(\mathbf{y}) = \neg f(\mathbf{x})$ and there exists some $S' \subseteq \overline{S}$ that is a sufficient reason of $\langle f, \mathbf{y} \rangle$.*

Proof. Given that S is sufficient for $\langle f, \mathbf{x} \rangle$, it follows that:

$$\forall (\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})] \tag{8}$$

By contradiction, let us assume that there exists some $\mathbf{y} \in \mathbb{F}$ for which $f(\mathbf{y}) = \neg f(\mathbf{x})$ and there exists some $S' \subseteq \overline{S}$ that is a sufficient reason of $\langle f, \mathbf{y} \rangle$. Since $S' \subseteq \overline{S}$ is a sufficient reason for $\langle f, \mathbf{y} \rangle$, this also implies that $\overline{S}$ is sufficient for $\langle f, \mathbf{y} \rangle$. In other words, the following condition holds:

$$\exists (\mathbf{y} \in \mathbb{F}), \; \forall (\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{y}_{\bar{S}}; \mathbf{z}_S) = f(\mathbf{y}) \neq f(\mathbf{x})] \tag{9}$$

Given that Equation 9 is valid for any $\mathbf{z} \in \mathbb{F}$, it is, consequently, applicable specifically to $\mathbf{x} \in \mathbb{F}$. In other words:

$$\exists (\mathbf{y} \in \mathbb{F}) \quad [f(\mathbf{y}_{\bar{S}}; \mathbf{x}_S) = f(\mathbf{y}) \neq f(\mathbf{x})] \tag{10}$$

This is inconsistent with the assertion that S is sufficient for $\langle f, \mathbf{x} \rangle$.

Lemma 3 *For a non-trivial function f, if S is a sufficient reason of $\langle f, \mathbf{x} \rangle$ then any $S' \subseteq \overline{S}$ is not a global sufficient reason of f.*

Proof. Given that S serves as a sufficient reason for $\langle f, \mathbf{x} \rangle$, it follows from Lemma 2 that there does not exist any $\mathbf{y} \in \mathbb{F}$ for which $f(\mathbf{y}) = \neg f(\mathbf{x})$ and $\overline{S}$ is sufficient for $\langle f, \mathbf{y} \rangle$. Consequently, if there indeed exists some $\mathbf{y} \in \mathbb{F}$ for which $\overline{S}$ serves as a sufficient reason for $\langle f, \mathbf{y} \rangle$, it necessarily follows that $f(\mathbf{x}) = f(\mathbf{y})$.

Let us, by contradiction, assume the existence of some $S' \subseteq \overline{S}$ that serves as a global sufficient reason of f. This implication further entails that $\overline{S}$ is also a global sufficient reason for f. Consequently, from the definition of global sufficient reasons, $\overline{S}$ is also a local sufficient reason for $\langle f, \mathbf{y} \rangle$ for any $\mathbf{y} \in \mathbb{F}$. Given the property highlighted earlier, it holds that for any $\mathbf{y} \in \mathbb{F}$, we have $f(\mathbf{y}) = f(\mathbf{x})$, which stands in contradiction to the premise that f is non-trivial.

We are now in a position to prove the first part of proposition 2:

Lemma 4 *If S_1 and S_2 are two global sufficient reasons of some non-trivial f, then $S_1 \cap S_2 \neq \emptyset$.*

Proof. Let us assume, to the contrary, that $S_1 \cap S_2 = \emptyset$. Hence, it follows that $S_1 \subseteq \overline{S_2}$. Given that S_2 is a global sufficient reason for f, it naturally follows that it is also a local sufficient reason for some $\langle f, \mathbf{x} \rangle$. Since S_2 is a local sufficient reason for some $\langle f, \mathbf{x} \rangle$, Lemma 3 determines that there does not exist any $S' \subseteq \overline{S_2}$ that can be a global sufficient reason for f. This is in direct contradiction with our earlier inference that $S_1 \subseteq \overline{S_2}$ is a global sufficient reason for f.

We now can proceed to prove the second part of proposition 2:

Lemma 5 *If S_1 and S_2 are global sufficient reasons of some non-trivial f ,then $S = S_1 \cap S_2$ is a global sufficient reason of f.*

Proof. First, from Lemma 4, it holds that $S \neq \emptyset$. In instances where either $S_1 \subseteq S_2$ or $S_2 \subseteq S_1$, the claim is straightforwardly true. Therefore, our remaining task is to prove the claim for a non-empty set S with the conditions $S \subsetneq S_1$ and $S \subsetneq S_2$.

Consider an arbitrary vector $\mathbf{x} \in \mathbb{F}$. Our aim is to prove that S is a local sufficient reason with respect to $\langle f, \mathbf{x} \rangle$. Should this hold true for an arbitrary $\mathbf{x}$, it follows that S constitutes a global sufficient reason of f.

Given that S_1 and S_2 are global sufficient reasons, it holds that:

$$\forall (\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_{S_1}; \mathbf{z}_{\bar{S}_1}) = f(\mathbf{x}) = f(\mathbf{x}_{S_2}; \mathbf{z}_{\bar{S}_2})] \tag{11}$$

To demonstrate that S is a local sufficient reason for $\langle f, \mathbf{x} \rangle$, let us assume, by contradiction, that it is not. Therefore, it satisfies that:

$$\begin{gathered} \exists (\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) \neq f(\mathbf{x})] \iff \\ \exists (\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{S_2 \setminus S}; \mathbf{z}_{\bar{S}_2}) \neq f(\mathbf{x})] \end{gathered} \tag{12}$$

Recall that S_2 is a global sufficient reason of f. Thus, assigning the features of S to the corresponding values in $\mathbf{x}$ determines the prediction. This also implies that fixing the features of S to the corresponding values $\mathbf{x}$ and assigning those of $S_2 \setminus S$ to the specific values of $\mathbf{z}$ from equation 12, determines that the prediction remains the same (which in this case is *not* the value $f(\mathbf{x})$). Formally put:

$$\forall (\mathbf{z}' \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{S_2 \setminus S}; \mathbf{z}'_{\bar{S}_2}) = f(\mathbf{x}_S; \mathbf{z}_{S_2 \setminus S}; \mathbf{z}_{\bar{S}_2}) \neq f(\mathbf{x})] \tag{13}$$

Let us define the set $S' = \{1, \ldots, n\} \setminus \{S_1 \cup S_2\}$. We now can equivalently express equation 13 as:

$$\forall (\mathbf{z}' \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{S_2 \setminus S}; \mathbf{z}'_{S_1 \setminus S}; \mathbf{z}'_{S'}) \neq f(\mathbf{x})] \tag{14}$$

But we know that S_1 is also a global sufficient reason and hence fixing the values of S_1 to $\mathbf{x}$ determines that the prediction is $f(\mathbf{x})$. Particularly, fixing the values of S_1 to $\mathbf{x}$ and the values of $S_2 \setminus S$ to $\mathbf{z}$ (from equation 12) still determines that the prediction is always $f(\mathbf{x})$.

$$\forall (\mathbf{z}' \in \mathbb{F}). \quad [f(\mathbf{x}_{S_1}; \mathbf{z}_{S_2 \setminus S}; \mathbf{z}'_{S' \cup S}) = f(\mathbf{x})] \tag{15}$$

Since the preceding statement remains valid for any partial assignment of the features in $S' \cup S$, we can consider a specific assignment where the features in S are set to their respective values in $\mathbf{x}$; thus, it is established that:

$$\forall (\mathbf{z}' \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{x}_{S_1}; \mathbf{z}_{S_2 \setminus S}; \mathbf{z}'_{S'}) = f(\mathbf{x})] \tag{16}$$

This particularly implies that:

$$\exists (\mathbf{z}' \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{S_2 \setminus S}; \mathbf{x}_{S_1 \setminus S}; \mathbf{z}'_{S'}) = f(\mathbf{x})] \tag{17}$$

which contradicts Equation 14.

Theorem 4 *There exists one unique subset-minimal global sufficient reason of f.*

Proof. First, for the scenario in which f is trivial (always outputs 1 or always outputs 0) it holds that any subset S is a global sufficient reason. Therefore, $S = \emptyset$ is a unique subset-minimal global sufficient reason. Let us now consider a non-trivial function f. Let us assume, by contradiction, that two distinct subset minimal global sufficient reasons of f exist: $S_1 \neq S_2$. Since S_1 and S_2 are subset minimal, it clearly holds that $S_1 \not\subseteq S_2$ and $S_2 \not\subseteq S_1$. Moreover, from Proposition 2 it can be asserted that S_1 and S_2 are not disjoint, i.e., $S_1 \cap S_2 \neq \emptyset$. Now, we can use Proposition 5, and conclude that $S = S_1 \cap S_2$ is also a global sufficient reason of f. This clearly contradicts the subset minimality of S_1 and S_2.

Proposition 3 *For any possible ordering of features in line 2 of Algorithm 2, Algorithm 2 converges to the same global sufficient reason.*

Since Algorithm 2 converges to a subset-minimal global sufficient reason, and there is only one unique subset-minimal global sufficient reason (Theorem 4), then iterating over any ordering of features in line 2 of Algorithm 2 will converge to the same subset. We note that the convergence of Algorithm 2 to a subset-minimal sufficient reason stems from the hereditary property of sufficient reasons, applicable to both local and global contexts. Specifically, if $S \subseteq \{1, \ldots, n\}$ is a (local/global) sufficient reason, then any subset $S' \subseteq \{1, \ldots, n\}$ that contains S (i.e., $S \subseteq S'$) will also be a (local/global) sufficient reason.

Proposition 4 *Let S be the unique subset minimal global sufficient reason of f. For every i, $i \in S$ if and only if i is locally necessary for some $\langle f, \boldsymbol{x} \rangle$, and $i \in \overline{S}$ if and only if i is globally redundant for f.*

We begin by proving the first part of the claim:

Lemma 6 *S is a unique subset-minimal global sufficient reason of f if and only if for every feature $i \in S$ it holds that i is locally necessary for some $\langle f, \boldsymbol{x} \rangle$.*

Proof. For the first direction, assume i is necessary for some $\langle f, \mathbf{x} \rangle$. Then, from Theorem 1, it holds that $\{i\}$ is contrastive for some $\langle f, \mathbf{x} \rangle$. Furthermore, the first duality theorem (Theorem 2), implies that each local contrastive reason intersects each global sufficient reason. Hence, we conclude that $i \in S$, for every global minimal sufficient reason S.

For the second direction, suppose that S is a unique subset-minimal global sufficient reason of f. Let there be some $i \in S$. Since S is *unique*, then $\{1, \ldots, n\} \setminus \{i\}$ is necessarily *not* a global sufficient reason. If this was so, then there would exist some subset $S' \subseteq \{1, \ldots, n\} \setminus \{i\}$ that is a subset-minimal global sufficient reason, contradicting the uniqueness of S.

Since $\{1, \ldots n\} \setminus \{i\}$ is not a global sufficient reason, there exist some $\mathbf{x}', \mathbf{z}' \in \mathbb{F}$ such that:

$$f(\mathbf{x}'_{\{1,\ldots n\} \setminus \{i\}}; \mathbf{z}'_{\{i\}}) \neq f(\mathbf{x}') \tag{18}$$

Thus, $\{i\}$ serves as a contrastive reason for $\langle f, \mathbf{x}' \rangle$ and from Theorem 1 we can infer that i is necessary with respect to $\langle f, \mathbf{x}' \rangle$.

For the second part of the claim, we prove the following Lemma:

Lemma 7 *Let S be the unique subset-minimal global sufficient reason of f. Then i is globally redundant if and only if $i \in \overline{S}$.*

Proof. For the first direction, let us assume that i is globally redundant and assume, by contradiction, that $i \in S$. Given that i is globally redundant for f then it holds that for any $\mathbf{x} \in \mathbb{F}$: $\text{suff}(f, \mathbf{x}, S) = 1 \rightarrow \text{suff}(f, \mathbf{x}, S \setminus \{i\}) = 1$. Hence, $S \setminus \{i\}$ is also a global sufficient reason of S, contradicting the subset-minimality of S.

For the second direction, assume that $i \in \overline{S}$. From Lemma 6, it holds that i is not locally necessary for any $\langle f, \mathbf{x} \rangle$. In other words, there does not exist any $\mathbf{x} \in \mathbb{F}$ for which $\text{suff}(f, \mathbf{x}, S) = 1 \rightarrow \text{suff}(f, \mathbf{x}, S \setminus \{i\}) = 0$. This implies that for any $\mathbf{x} \in \mathbb{F}$ it satisfies that $\text{suff}(f, \mathbf{x}, S) = 1 \rightarrow \text{suff}(f, \mathbf{x}, S \setminus \{i\}) = 1$, i.e., that i is globally redundant with respect to f.

Proposition 5 *Any feature i is either locally necessary for some $\langle f, \boldsymbol{x} \rangle$ or globally redundant for f.*

Building upon Proposition 4, we can discern that the unique subset minimal sufficient reason S of f categorizes all features into two distinct categories: those that are locally necessary for some $\langle f, \mathbf{x} \rangle$ and those that are globally redundant for f.

E. Proof of proposition 6

Proposition 6 *For Perceptrons, solving G-CSR and G-MSR is coNP-Complete, while solving CSR and MSR can be done in polynomial time.*

As mentioned in Table 1, the complexity of *CSR* and *MSR* for Perceptrons are drawn from the work of (Barceló et al., 2020) (similar proofs appear in (Marques-Silva et al., 2020)). We now move to prove all other complexity classes:

Lemma 8 *G-CSR is coNP-Complete for Perceptrons.*

Proof. Membership is straightforward since we can simply guess some $\mathbf{x}, \mathbf{z} \in \mathbb{F}$ and validate whether it satisfies that $f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) \neq f(\mathbf{x})$. If so, $\langle f, S \rangle \notin$ *G-CSR*.

We now will proceed to prove that *G-CSR* is also coNP-hard, We first briefly describe how the problem of (local) *CSR* can be solved in polynomial time for perceptrons, as proven by (Barceló et al., 2020) (a similar proof for a general linear classifier appears in (Marques-Silva et al., 2020)). This will give better intuition for the hardness reduction in the global setting. Given some $\langle f, \mathbf{x}, S \rangle$, recall that a Perceptron f is defined as $f = \langle \mathbf{w}, b \rangle$, where $\mathbf{w}$ is the weight vector and b is the bias term. Therefore, it is possible to obtain the exact value of $\sum_{i \in S} \mathbf{x}_i \cdot \mathbf{w}_i$.

Then, for the remaining features in $\overline{S}$, one can linearly determine the $\mathbf{y}$ assignments corresponding to the *maximal* and *minimal* values of $\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i$. The maximal value is obtained by setting $\mathbf{y}_i := 1$ when $\mathbf{w}_i \geq 0$ and $\mathbf{y}_i := 0$ when $\mathbf{w}_i \leq 0$. The minimal value is obtained respectively (setting $\mathbf{y}_i := 1$ when $\mathbf{w}_i < 0$ and $\mathbf{y}_i := 0$ when $\mathbf{w}_i \geq 0$). We are now able to compute the full range of potential values that may be realized by assigning the values of S to $\mathbf{x}$. It is hence straightforward that S is a (local) sufficient reason for $\langle f, \mathbf{x} \rangle$ if and only if this entire range is always positive or always negative. This can be determined by checking whether both the minimal possible value and maximal possible value are both positive or negative which is equivalent to checking whether the maximal possible value is non-negative or that the minimal possible value is positive. Formally put:

$$\begin{aligned} &\sum_{i \in S} \mathbf{x}_i \cdot \mathbf{w}_i + max\{\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i + b \mid \mathbf{y} \in \mathbb{F}\} \leq 0 \ \vee \\ &\quad \sum_{i \in S} \mathbf{x}_i \cdot \mathbf{w}_i + min\{\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i + b \mid \mathbf{y} \in \mathbb{F}\} > 0 \end{aligned} \tag{19}$$

This can clearly be determined in linear time using the computation method outlined above. Note that we require a strict inequation on the second term since we assumed w.l.o.g. that a zero weighted term is classified as 0 (the negative weighted class) and not 1 (the positive weighted class).

Now, for the global setting, we notice that $max\{\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i + b \mid \mathbf{y} \in \mathbb{F}\}$ and $min\{\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i + b \mid \mathbf{y} \in \mathbb{F}\}$ can still be computed in the same manner as above. However, one must verify that equation 19 is satisfied for *every* possible value $\mathbf{x}$. This, in turn, carries implications for the associated complexity. We show, indeed, that *G-CSR* for perceptrons is coNP-hard.

We reduce *G-CSR* for Perceptrons from $\overline{SSP}$, known to be coNP-Complete. *SSP* (subset-sum-problem) is a classic NP-Complete problem which is defined as follows:

> **SSP (Subset Sum Problem)**:
> **Input**: $\langle (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n), T \rangle$, where $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ is a set of *positive integers* and T, is the target integer.
> **Output**: *Yes*, if there exists a subset $S' \subseteq \{1, 2, \ldots, n\}$ such that $\sum_{i \in S'} \mathbf{z}_i = T$, and *No* otherwise.

For the case of $\overline{SSP}$, the language decides whether there does not exist a subset of features $S' \subseteq (1, 2, \ldots, n)$ for which $\sum_{i \in S'} \mathbf{z}_i = T$, i.e., for *all* subsets it holds that $\sum_{i \in S'} \mathbf{z}_i \neq T$.

We reduce *G-CSR* for Perceptrons from $\overline{SSP}$. Given some $\langle (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n), T \rangle$ we construct a Perceptron $f := \langle \mathbf{w}, b \rangle$ such that $\mathbf{w} := (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n) \cdot (\mathbf{w}_{n+1})$ ($\mathbf{w}$ is of size $n+1$), where $\mathbf{w}_{n+1} := \frac{1}{2}$, and $b := -(T + \frac{1}{4})$. The reduction computes $\langle f, S := \{1, \ldots, n\} \rangle$.

Clearly, it holds that:

$$max\{\sum_{i\in\bar{S}}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = max\{\sum_{i\in\{1,\ldots,n+1\}\setminus\{1,\ldots,n\}}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = \\ max\{\sum_{i=n+1}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = max\{\frac{1}{2},0\} = \frac{1}{2} \tag{20}$$

and that:

$$min\{\sum_{i\in\bar{S}}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = min\{\sum_{i\in\{1,\ldots,n+1\}\setminus\{1,\ldots,n\}}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = \\ min\{\sum_{i=n+1}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = min\{\frac{1}{2},0\} = 0 \tag{21}$$

If $\langle(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n),T\rangle \in \overline{SSP}$, there does not exist a subset $S'\subseteq S=\{1,2,\ldots,n\}$ for which $\sum_{i\in S'}\mathbf{z}_i = T$, put differently — for any subset $S'\subseteq S=\{1,2,\ldots,n\}$ it holds that $\sum_{i\in S'}\mathbf{z}_i > T$ or $\sum_{i\in S'}\mathbf{z}_i < T$. But since the values in $(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n)$ are *positive integers* then it also holds that for any subset S' the following condition is met:

$$\begin{aligned}
&\forall S'\subseteq S \quad [\,[\sum_{i\in S'}\mathbf{z}_i > T+\frac{1}{4}] \;\vee [\sum_{i\in S'}\mathbf{z}_i < T-\frac{1}{4}]\,] \iff \\
&\forall S'\subseteq S \quad [\,[\sum_{i\in S'}\mathbf{w}_i > T+\frac{1}{4}] \;\vee [\sum_{i\in S'}\mathbf{w}_i < T-\frac{1}{4}]\,] \iff \\
&\forall S'\subseteq S \quad [\,[\sum_{i\in S'}\mathbf{w}_i\cdot\mathbf{1}_i + \sum_{i\in S\setminus S'}\mathbf{w}_i\cdot\mathbf{0}_i > T+\frac{1}{4}] \;\vee [\sum_{i\in S'}\mathbf{w}_i\cdot\mathbf{1}_i + \sum_{i\in S\setminus S'}\mathbf{w}_i\cdot\mathbf{0}_i < T-\frac{1}{4}]\,] \iff \\
&\forall \mathbf{x}\in\{0,1\}^n \quad [\,[\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i > T+\frac{1}{4}] \;\vee\; [\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i < T-\frac{1}{4}]\,] \iff \\
&\forall \mathbf{x}\in\{0,1\}^n \quad [\,[\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + b > 0] \;\vee\; [\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + b < -\frac{1}{2}]\,]
\end{aligned} \tag{22}$$

Since we know from equations 20 and 21 that:

$$[\,min\{\sum_{i\in\bar{S}}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = 0] \quad \wedge [\,max\{\sum_{i\in\bar{S}}\mathbf{y}_i\cdot\mathbf{w}_i \mid \mathbf{y}\in\mathbb{F}\} = \frac{1}{2}\,] \tag{23}$$

This, combined with the result from equation 22, implies that:

$$\begin{aligned}
&\forall \mathbf{x}\in\{0,1\}^n \quad [\,[\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + b > 0] \;\vee\; [\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + b < -\frac{1}{2}]\,] \iff \\
&\forall \mathbf{x}\in\{0,1\}^{n+1} \quad [\,[\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + \mathbf{x}_{n+1}\cdot\mathbf{w}_{n+1} + b > 0] \;\vee\; [\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + \mathbf{x}_{n+1}\cdot\mathbf{w}_{n+1} + b < 0]\,] \iff \\
&\forall \mathbf{x},\mathbf{y}\in\{0,1\}^{n+1} \quad [\,[\sum_{i\in S}\mathbf{x}_i\cdot\mathbf{w}_i + \sum_{i\in\overline{S}}\mathbf{y}_i\cdot\mathbf{w}_i + b > 0] \;\vee\; [\sum_{i\in S}\mathbf{x}_i\cdot w_i + \sum_{i\in\overline{S}}\mathbf{y}_i\cdot\mathbf{w}_i + b < 0]\,] \iff \\
&\forall \mathbf{x},\mathbf{y}\in\{0,1\}^{n+1} \quad [\,[f(\mathbf{x}_S;\mathbf{y}_{\bar{S}}) = f(\mathbf{x}_S;\neg\mathbf{y}_{\bar{S}}) = 0] \;\vee\; [f(\mathbf{x}_S;\mathbf{y}_{\bar{S}}) = f(\mathbf{x}_S;\neg\mathbf{y}_{\bar{S}}) = 1]\,]
\end{aligned} \tag{24}$$

This implies that for any $\mathbf{x}\in\mathbb{F}$, fixing the values of $S=\{1,\ldots,n\}$ always maintains either a positive value (classified to 1) or a negative value (classified to 0) for f over any possible $\mathbf{y}$, thus implying that S is a global sufficient reason of f and that $\langle f,S\rangle \in$ *G-CSR*.

If $\langle(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n),T\rangle \notin \overline{SSP}$, then there exists a subset $S'\subseteq S=\{1,2,\ldots,n\}$ for which $\sum_{s_i\in S'}\mathbf{z}_i = T$, implying that:

$$\exists S' \subseteq S \quad [T-\frac{1}{4}] < \sum_{i\in S'} \mathbf{z}_i < [T+\frac{1}{4}] \iff \exists S' \subseteq S \quad [T-\frac{1}{4}] < \sum_{i\in S'} \mathbf{w}_i < [T+\frac{1}{4}] \iff$$
$$\exists S' \subseteq S \quad [T-\frac{1}{4}] < \sum_{i\in S'} \mathbf{w}_i \cdot \mathbf{1}_i + \sum_{i\in S\setminus S'} \mathbf{w}_i \cdot \mathbf{0}_i < [T+\frac{1}{4}] \iff \tag{25}$$
$$\exists \mathbf{x} \in \{0,1\}^n \quad [T-\frac{1}{4}] < \sum_{i\in S} \mathbf{w}_i \cdot \mathbf{x}_i < [T+\frac{1}{4}] \iff \exists \mathbf{x} \in \{0,1\}^n \quad (-\frac{1}{2}) < \sum_{i\in S} \mathbf{w}_i \cdot \mathbf{x}_i + b < 0$$

From here it holds that:

$$[\ \exists \mathbf{x} \in \{0,1\}^n \quad 0 < \sum_{i\in S} \mathbf{w}_i \cdot \mathbf{x}_i + b + \mathbf{w}_{n+1} \cdot 1 < (\frac{1}{2})\ ] \ \wedge$$
$$[\ \exists \mathbf{x} \in \{0,1\}^n \quad (-\frac{1}{2}) < \sum_{i\in S} \mathbf{w}_i \cdot \mathbf{x}_i + b + \mathbf{w}_{n+1} \cdot 0 < 0\ ] \tag{26}$$

This implies that:

$$\exists \mathbf{x}, \mathbf{y}, \mathbf{y}' \in \{0,1\}^{n+1} \ [\ 0 < \sum_{i\in S} \mathbf{w}_i \cdot \mathbf{x}_i + \mathbf{w}_{n+1} \cdot \mathbf{y}_{n+1} + b < (\frac{1}{2})\ ] \wedge$$
$$[\ -(\frac{1}{2}) < \sum_{i\in S} \mathbf{w}_i \cdot \mathbf{x}_i + \mathbf{w}_{n+1} \cdot \mathbf{y}'_{n+1} + b < 0] \tag{27}$$

This is equivalent to saying there exist features $\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \{0,1\}^{n+1}$ such that the output of the perceptron f for the input $(\mathbf{x}_S; \mathbf{y}_{\bar{S}})$ always lies in the range from 0 to $\frac{1}{2}$ (thereby being positive), and the output for the input $(\mathbf{x}_S; \mathbf{y}'_{\bar{S}})$ ranges from $-(\frac{1}{2})$ to 0 (thereby being negative). Hence:

$$\exists \mathbf{x}, \mathbf{y}, \mathbf{y}' \in \{0,1\}^{n+1} \ [\ f(\mathbf{x}_S; \mathbf{y}_{\bar{S}}) = 1] \ \wedge [\ f(\mathbf{x}_S; \mathbf{y}'_{\bar{S}}) = 0] \tag{28}$$

Thus, there exists some $\mathbf{x}$ for which S is not a sufficient reason with respect to $\langle f, \mathbf{x}\rangle$, indicating that S is not a global sufficient reason for f and $\langle f, S\rangle \notin$ *G-CSR*. This concludes the reduction.

Hardness results for Perceptrons, clearly indicate coNP-hardness for MLPs.

Lemma 9 *G-MSR is coNP-Complete for Perceptrons.*

Proof. **Membership.** Membership is derived from the fact that one can guess some $\mathbf{x}^1, \dots \mathbf{x}^n \in \mathbb{F}$ and $\mathbf{z}^1, \dots, \mathbf{z}^n \in \mathbb{F}$. We then can validate for every feature $i \in (1, \dots, n)$ whether: $f(\mathbf{x}^i_{\{1,\dots,n\}\setminus\{i\}}; \mathbf{z}^i_{\{i\}}) \neq f(\mathbf{x}^i)$. This will imply that $\{i\}$ is contrastive with respect to $\langle f, \mathbf{x}^i\rangle$ and from Theorem 1, i is necessary with respect to $\langle f, \mathbf{x}^i\rangle$. Now, from Proposition 4 it holds that i is contained in the unique global subset minimal sufficient reason of f if and only if i is necessary with respect to some $\langle f, \mathbf{x}\rangle$. Therefore, it is possible to validate whether $\langle f, k\rangle \notin$ *G-MSR* using a certificate that checks whether the number of features that satisfy: $f(\mathbf{x}^i_{\{1,\dots,n\}\setminus\{i\}}; \mathbf{z}^i_{\{i\}}) \neq f(\mathbf{x}^i)$ is larger than k.

Hardness. For hardness, we perform a similar reduction to the one performed for *G-CSR* for Perceptrons and reduce *G-MSR* for Perceptrons from $\overline{SSP}$. Given some $\langle(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n), T\rangle$, construct a Perceptron $f := \langle \mathbf{w}, b\rangle$ where we define $\mathbf{w} := (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n) \cdot (\mathbf{w}_{n+1})$ (i.e., $\mathbf{w}$ is of size $n+1$), where $\mathbf{w}_{n+1} := \frac{1}{2}$, and $b := -(T + \frac{1}{4})$. The reduction computes $\langle f, k := n\rangle$. We observe that this reduction is precisely the same as the one presented in Lemma 8, with the only difference being that we define $k := n$ rather than $S := \{1, \dots, n\}$. Therefore, we will base some of our assertions on the validity of the reduction from Lemma 8.

Consider that $\langle(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n), T\rangle \in \overline{SSP}$. Drawing upon the correctness of the reduction in Lemma 8, we can infer that $S = \{1, 2, \dots, n\}$ constitutes a global sufficient reason of f. To put it differently, a subset exists — trivially of size $k = n$ in this instance — that serves as a global sufficient reason of f. Consequently, $\langle f, k\rangle \in$ *G-MSR* for Perceptrons.

Assume $\langle(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n), T\rangle \notin \overline{SSP}$. We need to prove that there does not exist any global sufficient reason of f of size k or less. Since any subset containing a sufficient reason is a sufficient reason, it is enough to show that there does not exist any global sufficient reason of *exactly* size k. From the correctness of the reduction presented in Lemma 8 we indeed already know that $\{1, \ldots, n\}$, which corresponds to the assignments of $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ is *not* a sufficient reason in this case. However, we still need to prove that there does not exist any *other* sufficient reason of size k.

Let $j \neq n+1$ be some feature and let $S := \{1, 2, \ldots, n+1\} \setminus \{j\}$ be some subset of features. We will now prove that S is not a global sufficient reason for f. First, since any $\mathbf{z}_j$ in $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ is a positive integer, and since $\mathbf{w}_{n+1} = \frac{1}{2}$ is also positive, then it holds that:

$$
\begin{aligned}
max\{\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i \mid \mathbf{y} \in \mathbb{F}\} = max\{\mathbf{z}_j, 0\} = \mathbf{z}_j \ \wedge \\
min\{\sum_{i \in \bar{S}} \mathbf{y}_i \cdot \mathbf{w}_i \mid \mathbf{y} \in \mathbb{F}\} = min\{\mathbf{z}_j, 0\} = 0
\end{aligned} \tag{29}
$$

This implies that that S is a global sufficient reason of f iff:

$$
\forall \mathbf{x} \in \{0,1\}^{n+1} \quad [\sum_{i \in S} \mathbf{x}_i \cdot \mathbf{w}_i > T] \ \vee \ [\sum_{i \in S} \mathbf{x}_i \cdot \mathbf{w}_i \leq T - \mathbf{z}_j] \tag{30}
$$

Within Lemma 8 we have already determined that if $\langle(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n), T\rangle \notin \overline{SSP}$, then the following holds:

$$
\begin{aligned}
\exists \mathbf{x} \in \{0,1\}^n \quad \sum_{i \in \{1,\ldots,n\}} \mathbf{x}_i \cdot \mathbf{w}_i = T \iff \\
\exists \mathbf{x} \in \{0,1\}^n \quad \sum_{i \in \{1,\ldots,n\} \setminus \{j\}} \mathbf{x}_i \cdot \mathbf{w}_i = T - \mathbf{z}_j \iff \\
\exists \mathbf{x} \in \{0,1\}^n \quad \sum_{i \in \{1,\ldots,n\} \setminus \{j\}} \mathbf{x}_i \cdot \mathbf{w}_i + \mathbf{w}_{n+1} = T - \mathbf{z}_j + \frac{1}{2} \iff \\
\exists \mathbf{x} \in \{0,1\}^{n+1} \quad \sum_{i \in \{1,\ldots,n,n+1\} \setminus \{j\}} \mathbf{x}_i \cdot \mathbf{w}_i = T - \mathbf{z}_j + \frac{1}{2}
\end{aligned} \tag{31}
$$

Now, since T and $\mathbf{z}_j$ are positive *integers*, then from equation 31 it holds that:

$$
\begin{aligned}
\exists \mathbf{x} \in \{0,1\}^{n+1} \quad \sum_{i \in \{1,\ldots,n,n+1\} \setminus \{j\}} \mathbf{x}_i \cdot w_i = T - \mathbf{z}_j + \frac{1}{2} \Rightarrow \\
\exists \mathbf{x} \in \{0,1\}^{n+1} \quad T - \mathbf{z}_j < \sum_{i \in \{1,\ldots,n,n+1\} \setminus \{j\}} \mathbf{x}_i \cdot \mathbf{w}_i < T
\end{aligned} \tag{32}
$$

From equation 30, this implies that $\{1, \ldots, n, n+1\} \setminus \{j\}$ is *not* a global sufficient reason of f. Since there does not exist any j for which $\{1, \ldots, n, n+1\} \setminus \{j\}$ is a global sufficient reason of f and since we have already determined that $\{1, \ldots, n\}$ is not a global sufficient reason of f, we are left with that there does not exist any global sufficient reason of size k, concluding the reduction.

F. Proof of Proposition 7

Proposition 7 *For FBDDs and Perceptrons, FN and G-FN can be solved in polynomial time. However, for MLPs, FN can be solved in polynomial time, while solving G-FN is coNP-Complete*

As mentioned in Table 1, the complexity of *FN* for FBDDs is drawn from the work of (Huang et al., 2023). We now move to prove all other complexity classes:

Lemma 10 *G-FN can be solved in polynomial time for FBDDs.*

Proof. Let $\langle f, i\rangle$ be an instance. We describe the following polynomial algorithm: we enumerate pairs of leaf nodes (v, v') that correspond to the paths (α, α'). We denote by α_S the subset of nodes from α that correspond to the features of S. Given the pair (α, α') we check if α and α' "match" on all features from $\{1, \ldots, n\} \setminus \{i\}$ (more formally, there do not exist two nodes $v_\alpha \in \alpha_{\{1,\ldots,n\}\setminus\{i\}}$ and $v_{\alpha'} \in \alpha'_{\{1,\ldots,n\}\setminus\{i\}}$ with the same input feature j and with different *path* output edges). If we find two paths α and α' that (i) match on all features in $\{1, \ldots, n\} \setminus \{i\}$, (ii) *do not* match on feature i (i.e., have different path output edges for input feature i), and (iii) have *the same* label (both classified as True, or both classified as False) the algorithm returns "False" (i.e, i is *not* globally necessary with respect to f). If we do not encounter any such pair (v, v'), the algorithm returns "True".

Clearly, if the algorithm encounters two paths (α, α') that satisfy these three conditions, then it can be concluded that $\{i\}$ is *not* contrastive with respect to any assignment $\mathbf{x}$ associated with α and α'. From Theorem 1, this implies that i is not necessary with respect to the corresponding instances of $\langle f, \mathbf{x}\rangle$. However, if no such pair was encountered, then there does not exist an input $\mathbf{x} \in \mathbb{F}$ for which $\{i\}$ is not contrastive. It hence holds that $\{i\}$ is contrastive for any $\langle f, \mathbf{x}\rangle$ and Theorem 1 thereby implies that i is necessary with respect to any $\langle f, \mathbf{x}\rangle$.

Lemma 11 *G-FN can be solved in linear time for Perceptrons.*

Proof. Given some $\langle f, i\rangle$ such that $f := \langle \mathbf{w}, b\rangle$ is some Perceptron, we can perform a similar process to the one described under Lemma 8 and calculate: $max\{\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} y_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\}$ as well as: $min\{\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} \mathbf{y}_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\}$ in polynomial time. We now simply need to check whether there exists any instance $\mathbf{x} \in \mathbb{F}$ for which:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w}_i + max\{\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} \mathbf{y}_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\} \leq 0 \ \vee \\ \mathbf{x}_i \cdot \mathbf{w}_i + min\{\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} \mathbf{y}_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\} > 0 \end{aligned} \tag{33}$$

This condition can obviously be validated in polynomial time since there are only two possible relevant scenarios ($\mathbf{x}_i = 1$ or $\mathbf{x}_i = 0$). If this condition holds for one of the two possibilities then there exists an instance $\mathbf{x} \in \mathbb{F}$ for which $\{1, \ldots, n\} \setminus \{i\}$ is a local sufficient reason of $\langle f, \mathbf{x}\rangle$. It thereby holds that $\{i\}$ is *not* a contrastive reason of $\langle f, \mathbf{x}\rangle$. Hence, we can use Theorem 1, and conclude that i is not necessary with respect to $\langle f, \mathbf{x}\rangle$, thus implying that i is also not globally necessary. Should equation 33 not hold, it follows that for any $\mathbf{x} \in \mathbb{F}$ the set $\{1, \ldots, n\} \setminus \{i\}$ does *not* constitute a local sufficient reason of $\langle f, \mathbf{x}\rangle$. This conveys that $\{i\}$ is a local contrastive reason for *any* $\langle f, \mathbf{x}\rangle$. Theorem 1 further establishes that i is necessary for any $\langle f, \mathbf{x}\rangle$, and hence i is consequently globally necessary.

Lemma 12 *FN is in PTIME for Perceptrons and MLPs*

Building upon the correctness of Theorem 1, we can deduce that determining the necessity of feature i in relation to $\langle f, \mathbf{x}\rangle$ aligns with verifying if $\{i\}$ serves as a contrastive reason for $\langle f, \mathbf{x}\rangle$. For both MLPs and Perceptrons, it is possible to compute both $f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \mathbf{1}_{\{i\}})$ and $f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \mathbf{0}_{\{i\}})$ and validate whether:

$$f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \mathbf{1}_{\{i\}}) \neq f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \mathbf{0}_{\{i\}}) \tag{34}$$

The given condition is satisfied if, and only if, $\{i\}$ is contrastive with respect to $\langle f, \mathbf{x}\rangle$, thereby ascertaining whether i is necessary in relation to $\langle f, \mathbf{x}\rangle$.

Lemma 13 *G-FN is coNP-Complete for MLPs.*

Proof. To obtain membership, given a feature i that we aim to verify as globally necessary with respect to f, we can guess an instance $\mathbf{x} \in \mathbb{F}$ and determine whether:

$$f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \neg\mathbf{x}_{\{i\}}) = f(\mathbf{x}) \tag{35}$$

In other words, we wish to validate whether fixing all features in $\{1, \ldots, n\} \setminus \{i\}$ to their values in $\mathbf{x}$, and negating only the value of feature i (to be $\neg\mathbf{x}_i$) changes the prediction of $f(\mathbf{x})$. Clearly, this holds if and only if $\{i\}$ is *not* a contrastive reason for $\langle f, \mathbf{x}\rangle$ and from Theorem 1 this holds if and only if i is *not* necessary with respect to $\langle f, \mathbf{x}\rangle$. Put differently, there exists some $\mathbf{x} \in \mathbb{F}$ for which i is not necessary with respect to $\langle f, \mathbf{x}\rangle$. This implies that $\langle f, i\rangle \notin$ *G-FN*.

For Hardness, we will make use of the following Lemma whose proof appears in the work of (Barceló et al., 2020).

Lemma 14 *If we have a Boolean circuit B, we can create an MLP f_B in polynomial time that represents an equivalent Boolean function with respect to B.*

We now prove hardness by reducing from *TAUT*, a well-known coNP-Complete problem which is defined as follows:

TAUT (Tautology):
Input: A boolean formula ψ
Output: *Yes*, if ψ is a tautology and *No* otherwise.

Given some $\langle\psi\rangle$ with variables: $x_1, \ldots x_n$ we introduce a new variable x_{n+1} and construct a new boolean formula:

$$\psi' := \psi \wedge x_{n+1} \tag{36}$$

We then can use Lemma 14 to transform it to an MLP f and construct $\langle f, i := n+1\rangle$.

If $\langle\psi\rangle \in$ *TAUT* then it holds that ψ is always True, and hence:

$$f(\mathbf{x}_{\{1,\ldots,n\}}; \mathbf{1}_{\{n+1\}}) = 1 \quad \wedge \quad f(\mathbf{x}_{\{1,\ldots,n\}}; \mathbf{0}_{\{n+1\}}) = 0 \tag{37}$$

where $\mathbf{1}_{\{n+1\}}$ and $\mathbf{0}_{\{n+1\}}$ denote that feature $n+1$ is set to either 1 or 0.

Hence, for any value $\mathbf{x} \in \mathbb{F}$ we can find a corresponding instance $\mathbf{z} \in \mathbb{F}$ such that:

$$f(\mathbf{x}_{\{1,\ldots,n\}}; \mathbf{z}_{\{n+1\}}) \neq f(\mathbf{x}) \tag{38}$$

This implies that the subset $\{n+1\}$ is contrastive with respect to any $\langle f, \mathbf{x}\rangle$ and from theorem 1, feature $n+1$ is necessary with respect to any $\langle f, \mathbf{x}\rangle$. Thus, it satisfies that feature $n+1$ is globally necessary with respect to f and consequently, $\langle f, i\rangle \in$ *G-FN*.

Let us now consider the scenario where $\langle\psi\rangle \notin$ *TAUT*. Under this assumption, it follows that there exists a False assignment for $\langle x_1, \ldots, x_n\rangle$, rendering ψ' false irrespective of the assignment to $\mathbf{x}_{n+1}$. To put it differently, this scenario satisfies the following condition:

$$f(\mathbf{x}_{\{1,\ldots,n\}}; \mathbf{1}_{\{n+1\}}) = 0 \quad \wedge \quad f(\mathbf{x}_{\{1,\ldots,n\}}; \mathbf{0}_{\{n+1\}}) = 0 \tag{39}$$

Thus, we can take an arbitrary vector $\mathbf{x}$ and set some other arbitrary vector $\mathbf{z}$ to be equal to $\mathbf{x}$ on the first n features and negated on feature $n+1$. Both of these vectors will be labeled to class 0, hence:

$$\exists \mathbf{z}, \mathbf{x} \in \mathbb{F} \quad f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \mathbf{z}_{\{n+1\}}) = f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}}; \neg\mathbf{z}_{\{n+1\}}) = 0 = f(\mathbf{x}) \tag{40}$$

We can thus conclude that $\{n+1\}$ is *not* a contrastive reason of $\langle f, \mathbf{x}\rangle$ and from theorem 1, this implies that $n+1$ is not necessary with respect to $\langle f, \mathbf{x}\rangle$. Particularly, $n+1$ is not globally necessary, consequently implying that $\langle f, i\rangle \notin$ *G-FN*.

G. Proof of Proposition 8

Proposition 8 *For FBDDs, G-CSR and G-MSR can be solved in polynomial time, while MSR is NP-Complete. Moreover, in MLPs, solving G-CSR and G-MSR is coNP-Complete, while solving MSR is Σ_2^P-Complete.*

As mentioned in Table 1, the complexity of *MSR* for FBDDs and *MSR* for MLPs are drawn from the work of (Barceló et al., 2020). We now move to prove all other complexity classes:

Lemma 15 *G-CSR can be solved in polynomial time for FBDDs.*

Proof. Let $\langle f, S\rangle$ be an instance. We describe the following polynomial algorithm: We enumerate pairs of leaf nodes (v, v') that correspond to the paths (α, α'). Let us denote by α_S the subset of nodes from α that correspond to the features of S. Given the pair (α, α'), the algorithm checks if α and α' "match" on all features from S (more formally, there do not exist two nodes $v_\alpha \in \alpha_S$ and $v_{\alpha'} \in \alpha'_S$ with the same input feature i and with different path output edges). If we find two paths α and α' that match all features in S, and that have different labels (one classified as True and the other: False) the algorithm returns "False" (i.e., S is not a global sufficient reason of f). If we do not encounter any such pair (v, v'), the algorithm returns True.

Lemma 16 *G-MSR is in PTIME for FBDDs.*

Proof. Since *G-CSR* is in PTIME for FBDDs, we can use Proposition 3 which states that algorithm 2 always converges to the unique global cardinally minimal sufficient reason after a linear number of calls checking whether $\text{suff}(f, S \setminus \{i\}) = 1$. Each one of these calls can be performed in polynomial time (since *G-CSR* is polynomial for FBDDs), so hence using algorithm 2, we can obtain the unique global cardinally minimal sufficient reason of f in polynomial time, and return True if its size is smaller or equal to k, and False otherwise.

Lemma 17 *G-CSR is coNP-Complete for MLPs.*

Proof. Membership is straightforward and is obtained since we can guess some $\mathbf{x}, \mathbf{z} \in \mathbb{F}$ and validate whether it satisfies that $f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) \neq f(\mathbf{x})$. If so, $\langle f, S\rangle \notin$ *G-CSR*.

Given our forthcoming proof that the *G-CSR* query for Perceptrons is coNP-Hard, it follows straightforwardly that the same is true for MLPs. Nevertheless, we show how hardness can also be proved particularly for MLPs via a reduction from the (local) *CSR* explainability query for MLPs.

Given the tuple $\langle f, \mathbf{x}, S\rangle$ we construct an MLP f' which satisfies the following conditions:

$$f'(\mathbf{y}) = \begin{cases} f(\mathbf{y}) & if \ \ (\mathbf{x}_S = \mathbf{y}_S) \\ 1 & if \ \ (\mathbf{x}_S \neq \mathbf{y}_S) \end{cases} \tag{41}$$

An MLP corresponding to this specification can be built by applying Lemma 14, which asserts that any boolean circuit can be represented as an equivalent MLP. Specifically, we can encode the condition $(\mathbf{x}_i \oplus \mathbf{y}_i)$ for each $i \in S$ and define:

$$\psi := \bigwedge_{i \in S} (\mathbf{x}_i \oplus y_i) \tag{42}$$

Using Lemma 14, we can convert ψ into an MLP g and subsequently integrate g with the input layer of the original MLP, f. This configuration results in the input layer of f receiving connections from two distinct MLPs, each generating different outputs. Further applying Lemma 14, we introduce an additional hidden layer that produces a single output representing the disjunction of the two preceding outputs. This newly constructed MLP corresponds to the structure of f'. To preserve the integrity of the MLP's architecture, we can implement zero-weights and zero-biases for any unconnected neuron connections.

If $\langle f, \mathbf{x}, S\rangle \in$ *CSR*, then it satisfies that:

$$\forall (\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})] \tag{43}$$

Given that $f'(\mathbf{y}) = f(\mathbf{y})$ holds for any input for which $\mathbf{x}_S = \mathbf{y}_S$, then it also satisfies that:

$$\begin{gathered} \forall (\mathbf{z} \in \mathbb{F}). \quad [f'(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f'(\mathbf{x})] \iff \\ \forall (\mathbf{x}, \mathbf{z} \in \mathbb{F}). \quad (\mathbf{x}_S = \mathbf{z}_S) \rightarrow [f'(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f'(\mathbf{x})] \end{gathered} \tag{44}$$

If $\mathbf{x}_S \neq \mathbf{y}_S$ then it consequently holds that $f'(\mathbf{y}) = 1$. This implies that:

$$\forall(\mathbf{x}, \mathbf{z} \in \mathbb{F}). \quad (\mathbf{x}_S \neq \mathbf{z}_S) \rightarrow [f'(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f'(\mathbf{x}) = 1] \tag{45}$$

Overall, we arrive at that:

$$\forall(\mathbf{x}, \mathbf{z} \in \mathbb{F}). \quad [f'(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f'(\mathbf{x})] \tag{46}$$

implying that $\langle f', S \rangle \in$ *G-CSR*.

If $\langle f, \mathbf{x}, S \rangle \notin$ *CSR*, then it satisfies that:

$$\exists(\mathbf{z} \in \mathbb{F}). \quad [f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) \neq f(\mathbf{x})] \tag{47}$$

Given that $f'(\mathbf{y}) = f(\mathbf{y})$, it follows that for any input satisfying $\mathbf{x}_S = \mathbf{y}_S$ the following condition is also met:

$$\exists(\mathbf{z} \in \mathbb{F}). \quad [f'(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) \neq f'(\mathbf{x})] \tag{48}$$

implying that:

$$\exists(\mathbf{x}, \mathbf{z} \in \mathbb{F}). \quad [f'(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) \neq f'(\mathbf{x})] \tag{49}$$

Thus, it holds that $\langle f', S \rangle \notin$ *G-CSR*.

Lemma 18 *G-MSR is coNP complete for MLPs.*

Both hardness and membership results trivially derive from those described for Perceptrons.

H. Proof of Proposition 9

Proposition 9 *For FBDDs, G-FR can be solved in polynomial time, while solving FR is coNP-Complete. Moreover, in MLPs, solving G-FR is coNP-Complete, while solving FR is* Π_2^P*-Complete.*

As mentioned in Table 1, the complexity of *MSR* for MLPs is drawn from the work of (Huang et al., 2021) which proved hardness for DNF classifiers (and this also holds for MLPs, from Lemma 14). Moreover, the complexity of *FR* for FBDDs is drawn from the work of (Huang et al., 2023). We now move to prove all other complexity classes:

Lemma 19 *G-FR can be solved in polynomial time for FBDDs.*

Proof. Let $\langle f, i \rangle$ be an instance. We describe the following polynomial algorithm: We enumerate pairs of leaf nodes (v, v') that correspond to the paths (α, α'). We denote by α_S the subset of nodes from α that correspond to the features of S. Given the pair (α, α') we check if α and α' "match" on all features from $\{1, \ldots, n\} \setminus \{i\}$ (more formally, there do not exist two nodes $v_\alpha \in \alpha_{\{1,\ldots,n\}\setminus\{i\}}$ and $v_{\alpha'} \in \alpha'_{\{1,\ldots,n\}\setminus\{i\}}$ with the same input feature j and with different output edges). If we find two paths α and α' that (i) match on all features in $\{1, \ldots, n\} \setminus \{i\}$, (ii) *do not* match on feature i (i.e., have different output edges), and (iii) have different labels (one is classified as True and the other: False) the algorithm returns "False" (i.e, i is *not* redundant with respect to f). If we do not encounter any such pair (v, v'), the algorithm returns "True".

Lemma 20 *G-FR is coNP-Complete for MLPs.*

Both hardness and membership proofs for Perceptrons (proved in the following section) also trivially hold for MLPs.

I. Proof of Proposition 10

Proposition 10 *For Perceptrons, solving FR and G-FR are both coNP-Complete.*

Lemma 21 *FR is coNP-Complete for Perceptrons*

Proof. **Membership.** We recall that validating whether a subset S is a local sufficient reason with respect to some $\langle f, \mathbf{x} \rangle$ can be done in polynomial time for Perceptrons, as was elaborated on in Lemma 8. This can be done by polynomially calculating both: $max\{\sum_{j \in \bar{S}} \mathbf{y}_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\}$ and $min\{\sum_{j \in \bar{S}} \mathbf{y}_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\}$ and then validating whether it holds that:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w}_i + max\{\sum_{j \in \bar{S}} \mathbf{y}_j \cdot \mathbf{w}_j + b \mid \mathbf{y} \in \mathbb{F}\} \leq 0 \ \vee \\ \mathbf{x}_i \cdot \mathbf{w}_i + max\{\sum_{j \in \bar{S}} \mathbf{y}_j \cdot w_j + b \mid \mathbf{y} \in \mathbb{F}\} > 0 \end{aligned} \tag{50}$$

Hence, membership in coNP holds since we can guess some subset $S \subseteq \{1, \ldots, n\}$ and polynomially validate whether it holds that:

$$\text{suff}(f, \mathbf{x}, S) = 1 \ \ \wedge \ \ \text{suff}(f, \mathbf{x}, S \setminus \{i\}) = 0 \tag{51}$$

If the following condition holds, then it satisfies that i is not redundant with respect to $\langle f, \mathbf{x} \rangle$ and hence $\langle f, i \rangle \notin$ *FR*.

Hardness. We reduce *FR* for Perceptrons from the subset sum problem (SSP), specifically from $\overline{SSP}$ which is coNP-Complete. Given some $\langle (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n), T \rangle$ construct a Perceptron $f := \langle \mathbf{w}, b \rangle$ where we set $\mathbf{w} := (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n) \cdot (\mathbf{w}_{n+1})$ ($\mathbf{w}$ is of size $n+1$), where $\mathbf{w}_{n+1} := \frac{1}{2}$, and $b := -(T + \frac{1}{4})$. The reduction computes $\langle f, \mathbf{x} := \mathbf{1}, i := n+1 \rangle$.

Assume that $\langle (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n), T \rangle \in \overline{SSP}$. This implies that there does not exist any subset $S \subseteq \{1, \ldots, n\}$ for which $\sum_{j \in S} \mathbf{z}_j = T$. Given that the values in $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ are *integers*, it consequently follows that there does not exist a subset S satisfying that:

$$T - \frac{1}{2} < \sum_{j \in S} \mathbf{z}_j < T + \frac{1}{2} \tag{52}$$

Consequently, it holds that there is no subset $S \subseteq \{1, \ldots, n\}$ for which:

$$\begin{aligned} T - \frac{1}{2} < \sum_{j \in S} \mathbf{w}_j \cdot 1 < T + \frac{1}{2} \iff \\ -\frac{3}{4} < \sum_{j \in S} \mathbf{w}_j \cdot 1 + b < \frac{1}{4} \end{aligned} \tag{53}$$

which is equivalent to saying that for any subset $S \subseteq \{1, \ldots, n\}$ it holds that:

$$[\sum_{j \in S} \mathbf{w}_j \cdot 1 + b < -\frac{3}{4}] \ \vee \ [\sum_{j \in S} \mathbf{w}_j \cdot 1 + b > \frac{1}{4}] \tag{54}$$

This implies that:

$$\begin{aligned} \forall S' \subseteq \{1, \ldots, n+1\} \ [\,[\sum_{j \in S} \mathbf{w}_j \cdot 1 + \mathbf{w}_{n+1} \cdot 0 + b < 0\,] \ \wedge \ [\sum_{j \in S} \mathbf{w}_j \cdot 1 + \mathbf{w}_{n+1} \cdot 1 + b < -\frac{1}{4}\,]\,] \ \vee \\ [\,[\sum_{j \in S} \mathbf{w}_j \cdot 1 + \mathbf{w}_{n+1} \cdot 0 + b > \frac{1}{4}\,]\,] \ \wedge \ [\sum_{j \in S} \mathbf{w}_j \cdot 1 + \mathbf{w}_{n+1} \cdot 1 + b > \frac{3}{4}\,]\,] \end{aligned} \tag{55}$$

Thus, if we set the values of all features $\{1,\ldots,n+1\}$ to 1, modifying the value of feature $n+1$ from 1 to 0 will not alter the classification (as f will continue to be either positive or negative). Equivalently:

$$\forall S' \subseteq \{1,\ldots,n\} \quad [\,[\, f(\mathbf{1}_{S'};\mathbf{0}_{\{n+1\}}) = 0 \;\wedge\; [\, f(\mathbf{1}_{S'};\mathbf{1}_{\{n+1\}}) = 0\,]\,] \vee \\ [\,[\, f(\mathbf{1}_{S'};\mathbf{0}_{\{n+1\}}) = 1 \;\wedge\; [\, f(\mathbf{1}_{S'};\mathbf{1}_{\{n+1\}}) = 1\,]\,] \tag{56}$$

In other words, it can be stated that:

$$\forall S' \subseteq \{1,\ldots,n,n+1\} \quad \text{suff}(f,\mathbf{1},S') = 1 \rightarrow \text{suff}(f,\mathbf{1},S' \setminus \{n+1\}) = 1 \tag{57}$$

Therefore, feature $n+1$ is redundant with respect to $\langle f,\mathbf{1}\rangle$, implying that $\langle f,\mathbf{1},i\rangle \in$ *FR*.

Let us assume that $\langle(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n),T\rangle \notin \overline{SSP}$. From this assumption, it follows that there exists a subset of features, $S \subseteq \{z_1,\ldots \mathbf{z}_n\}$ for which: $\sum_{j\in S}\mathbf{z}_j = T$. We can express this equivalently as:

$$T = \sum_{j\in S}\mathbf{z}_j \iff -\frac{1}{4} = \sum_{j\in S}\mathbf{w}_j + b \iff \\ [-\frac{1}{4} = \sum_{j\in S}\mathbf{w}_j + \mathbf{w}_{n+1}\cdot 0 + b] \;\wedge\; [\frac{1}{4} = \sum_{j\in S}\mathbf{w}_j + \mathbf{w}_{n+1}\cdot 1 + b] \tag{58}$$

We denote $S' := S \cup \{n+1\}$. Based on equation 58, we have that $f(\mathbf{1}_{S'};\mathbf{0}_{\bar{S}'}) > 0$. Moreover, given that all features in $\bar{S}'$ are positive integers, then if we add additional features to S' (making it larger) the classification will necessarily remain positive. More formally, it is also established that for any $S'' \subseteq \{1,\ldots,n+1\}$ for which $S' \subseteq S''$ the following holds: $f(\mathbf{1}_{S''};\mathbf{0}_{\bar{S}''}) > 0$. Hence, fixing the features of S' to value 1 maintains a positive value for f (and hence a 1 classification), and hence S' is sufficient with respect to $\langle f,\mathbf{1}\rangle$. Referring to equation 58, we observe that: $f(\mathbf{1}_{S'\setminus\{n+1\}};\mathbf{0}_{\bar{S}'\cup\{n+1\}}) = 0$. This implies that $S' \setminus \{n+1\}$ is *not* sufficient with respect to $\langle f,\mathbf{1}\rangle$. In other words, we can conclude that:

$$\exists S' \subseteq \{1,\ldots,n,n+1\} \quad \text{suff}(f,\mathbf{1},S') = 1 \wedge \text{suff}(f,\mathbf{1},S' \setminus \{n+1\}) = 0 \tag{59}$$

Consequently, feature $n+1$ is *not* redundant with respect to $\langle f,\mathbf{1}\rangle$, thus implying that $\langle f,\mathbf{x},i\rangle \notin$ *FR*.

Lemma 22 *G-FR is coNP-Complete for Perceptrons.*

Proof. Membership is established from the fact that one can guess some $\mathbf{x},\mathbf{z} \in \mathbb{F}$ and validate whether: $f(\mathbf{x}_{\{1,\ldots,n\}\setminus\{i\}};\mathbf{z}_{\{i\}}) \neq f(\mathbf{x})$. From Theorem 1, this condition holds if and only if i is necessary with respect to $\langle f,\mathbf{x}\rangle$. Furthermore, Proposition 5 establishes that this situation is equivalent to i being *not* globally redundant with respect to f, thereby implying $\langle f,i\rangle \notin$ *G-FR*.

Before proving hardness, we will make use of the following Lemma which is simply a refined version of Proposition 4:

Lemma 23 *S is a global sufficient reason of f iff for any $i \in \overline{S}$, i is globally redundant.*

Proof. S is a global sufficient reason of f if and only if there exists some $S' \subseteq S$ which is a subset minimal global sufficient reason of f. From Proposition 4, it holds that any feature $i \in \bar{S}'$ is globally redundant, and since $\bar{S} \subseteq \bar{S}'$, it satisfies that any feature $i \in \bar{S}$ is globally redundant.

We are now in a position to employ Lemma 23, from which we can discern that S qualifies as a global sufficient reason of f if and only if every $i \in \bar{S}$ is globally redundant. Consequently, we can leverage the reduction that was utilized for establishing the coNP-Hardness of *G-CSR* for Perceptrons, as detailed in Lemma 8.

In other words, we can reduce *G-FR* for Perceptrons from $\overline{SSP}$. Given some $\langle(\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n),T\rangle$ we can again construct a Perceptron $f := \langle\mathbf{w},b\rangle$ where $\mathbf{w} := (\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n)\cdot(\mathbf{w}_{n+1})$ ($\mathbf{w}$ is of size $n+1$), $\mathbf{w}_{n+1} := \frac{1}{2}$, and $b := -(T+\frac{1}{4})$. The reduction computes $\langle f,i := n+1\rangle$.

It has been established in Lemma 8 that $S = \{1, 2, \ldots, n\}$ serves as a global sufficient reason of f if and only if no subset $S' \subseteq S = \{1, 2, \ldots, n\}$ exists for which $\sum_{i \in S'} \mathbf{z}_i = T$. Moreover, due to Lemma 23, the set $S = \{1, 2, \ldots, n\}$ is a global sufficient reason of f if and only if any feature in $\bar{S}$ is globally redundant. However, $\bar{S}$ is in our case simply $\{n+1\}$. This leads to the conclusion that feature $n+1$ is globally redundant, thereby concluding the reduction.

J. Proof of Proposition 11

Proposition 11 *For FBDDs, both CC and G-CC can be solved in polynomial time. Moreover, for Perceptrons and MLPs, solving both CC or G-CC is $\#P$-Complete.*

Lemma 24 *G-CC is $\#P$-Complete for Perceptrons.*

For simplification, we follow common conventions (Barceló et al., 2020) and prove that the global counting procedure for: $C(S, f) = |\{\mathbf{x} \in \mathbb{F}, \mathbf{z} \in \{0,1\}^{|\overline{S}|}, f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})\}|$ is $\#P$-Complete, rather than $c(S, f)$. Clearly, it holds that: $C(S, f) = c(S, f) \cdot 2^{|\bar{S}|+n}$ and hence $c(S, f)$ and $C(S, f)$ are interchangeable. We observe that computing $2^{|\bar{S}|+n}$ can be executed in polynomial time because n, which denotes the number of input features, is given in *unary* (the full size of the input encoding is at least n).

Membership. Membership is straightforward since the sum: $|\{\mathbf{x} \in \mathbb{F}, \mathbf{z} \in \{0,1\}^{|\overline{S}|}, f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x})\}|$ is equivalent to the sum of certificates $(\mathbf{x}, \mathbf{z})$ satisfying:

$$\exists \mathbf{x} \in \mathbb{F}, \exists \mathbf{z} \in \{0,1\}^{|\overline{S}|}, f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = f(\mathbf{x}) \tag{60}$$

which is of course polynomially verifiable.

Hardness. We reduce from (local) *CC* of Perceptrons, which is $\#P$-Complete (Barceló et al., 2020). Given some $\langle f, S, \mathbf{x} \rangle$, where $f := \langle \mathbf{w}, b \rangle$ is a Perceptron, the reduction computes $f(\mathbf{x})$ and if $f(\mathbf{x}) = 1$ constructs $f' := \langle \mathbf{w}', b' \rangle$ such that $b' := b + \sum_{i \in S} (\mathbf{x}_i \cdot \mathbf{w}_i)$, and $\mathbf{w}' := (\mathbf{w}_{\bar{S}}, \delta)$, with $\delta := (\sum_{i \in \overline{S}} |w_i|) - b'$. $\mathbf{w}_{\bar{S}}$ denotes a partial assignment where all features of the subset $\bar{S}$ are drawn from the vector $\mathbf{w}$ (the vector $\mathbf{w}'$ is of size $|\bar{S}| + 1$). If $f(\mathbf{x}) = 0$, the reduction constructs $f' := \langle \mathbf{w}', b' \rangle$ with the same b' but with $\mathbf{w}' := (\mathbf{w}_{\bar{S}}, \delta')$, such that $\delta' := -(\sum_{i \in \overline{S}} |\mathbf{w}_i|) - b' - 1$.

For both reduction scenarios ($f(\mathbf{x}) = 1$ or $f(\mathbf{x}) = 0$) we will demonstrate that given the *global* completion count $C(\emptyset, f')$ we can determine the *local* completion count of $c(S, f, \mathbf{x})$ in polynomial time. We do this by proving the following Lemma:

Lemma 25 *Given the polynomial construction of f', it satisfies that:*

$$C(S, f, \boldsymbol{x}) = \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{61}$$

If this lemma is proven, then demonstrating the hardness of the global completion count for perceptrons becomes straightforward. This is because n, representing the number of input features, is given in unary. Consequently, computations such as $2^{O(n)}$ (and therefore $2^{2|\overline{S}|}$) can be performed in polynomial time. Additionally, since the *binary* representation of $[\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}]$ is of size $O(n)$ ($C(\emptyset, f')$ is bounded by 2^{2n}), computing the square root of this expression can also be achieved in polynomial time.

Proof of Lemma. We denote m and t as the *number* of assignments $\mathbf{x}' \in \{0,1\}^{|\overline{S}|+1}$, for which f' predicts 0 or 1. Formally:

$$m := \left| \left\{ \mathbf{x}' \in \{0,1\}^{|\overline{S}|+1} \;\middle|\; f'(\mathbf{x}') = 1 \right\} \right| \quad , \quad t := \left| \left\{ \mathbf{x}' \in \{0,1\}^{|\overline{S}|+1} \;\middle|\; f'(\mathbf{x}') = 0 \right\} \right| \tag{62}$$

Clearly, it holds that:

$$m + t = 2^{|\overline{S}|+1} \tag{63}$$

It also satisfies that:

$$C(S:=\emptyset, f') = |\{\mathbf{x}' \in \{0,1\}^{|\overline{S}|+1}, \mathbf{z} \in \{0,1\}^{|\overline{S}|+1}, f'(\mathbf{x}'_S;\mathbf{z}_{\bar{S}}) = f'(\mathbf{x}')\}| = \\ |\{\mathbf{x}' \in \{0,1\}^{|\overline{S}|+1}, \mathbf{z} \in \{0,1\}^{|\overline{S}|+1}, f'(\mathbf{z}) = f'(\mathbf{x}')\}| = m^2 + t^2 \tag{64}$$

This is true because there are precisely m^2 pairs $(\mathbf{x}', \mathbf{z})$ where both $f'(\mathbf{z}) = 1$ and $f'(\mathbf{x}') = 1$, and exactly t^2 pairs where $f'(\mathbf{z}) = 0$ and $f'(\mathbf{x}') = 0$. As a result of equations 63 and 64, it satisfies that:

$$C(\emptyset, f') = m^2 + (2^{|\bar{S}|+1} - m)^2 \tag{65}$$

This implies that the aforementioned values of m/t obey the following quadratic relation:

$$\begin{aligned} m/t &= \frac{-(-2^{|\overline{S}|+2}) \pm \sqrt{(-2^{|\overline{S}|+2})^2 - 4 \cdot 2 \cdot (2^{2|\overline{S}|+2} - C(\emptyset, f'))}}{2 \cdot 2} \\ &= \frac{2^{|\overline{S}|+2} \pm \sqrt{2^{2|\overline{S}|+4} - 8 \cdot (2^{2|\overline{S}|+2} - C(\emptyset, f'))}}{4} \\ &= 2^{|\overline{S}|} \pm \sqrt{2^{2|\overline{S}|} - (2^{2|\overline{S}|+1} - \frac{1}{2} \cdot C(\emptyset, f'))} \end{aligned}$$

Accordingly, m/t must obey the following condition:

$$m/t = 2^{|\overline{S}|} \pm \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{66}$$

We observe that one of the terms exceeds $2^{|\overline{S}|}$, while the other is less than $2^{|\overline{S}|}$. Therefore, to ascertain whether m or t corresponds to the first or second term, it suffices to compare their counts to $2^{|\overline{S}|}$. We will start by proving the first part of the Lemma. Specifically, to establish that when $f(\mathbf{x}) = 1$ the following condition is satisfied:

$$C(S, f, \mathbf{x}) = \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{67}$$

We will first prove that when $f(\mathbf{x}) = 1$, then there are at least $2^{|\bar{S}|}$ vectors $\mathbf{x}' \in \{0,1\}^{|\bar{S}|+1}$ for which $f'(\mathbf{x}') = 1$.

First, we assume that $\mathbf{x}'_{|\bar{S}|+1} = 1$ (the assignment for feature $|\bar{S}| + 1$ is 1). For all $\mathbf{x}' \in \{0,1\}^{|S|+1}$ such that $\mathbf{x}'_{|\bar{S}|+1} = 1$ it holds that:

$$\begin{aligned} \mathbf{w}' \cdot \mathbf{x}' + b' &= \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \mathbf{w}'_{|\overline{S}|+1} \cdot \mathbf{x}'_{|\overline{S}|+1} + b' &= \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \delta + b' &= \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \left((\sum_{i \in \overline{S}} |\mathbf{w}_i|) - b' \right) + b' &= \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \sum_{i \in \overline{S}} |\mathbf{w}_i| &\geq 0 \end{aligned} \tag{68}$$

Given that there are precisely $2^{|\overline{S}|}$ assignments where $\mathbf{x}'_{|\overline{S}|+1} = 1$, it can be inferred that there are at least $2^{|\overline{S}|}$ assignments for which $f'(\mathbf{x}') = 1$. Hence, the following condition holds:

$$m = 2^{|\overline{S}|} + \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{69}$$

The count in equation 69 corresponds to the number of positive assignments for which $f(\mathbf{x}') = 1$. As mentioned, out of these, there are $2^{|\overline{S}|}$ assignments where $\mathbf{x}'_{|S|+1} = 1$. Consequently, the exact number of assignments with $\mathbf{x}'_{|S|+1} = 0$ that satisfy that $f'(\mathbf{x}) = 1$ is exactly:

$$(2^{|\overline{S}|} + \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}}) - 2^{|\overline{S}|} = \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{70}$$

Furthermore, it holds that:

$$\begin{aligned} f'(\mathbf{x}'_{\bar{S}}; \mathbf{0}_{|\bar{S}|+1}) = step(\mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + 0 + b') = \\ step(\mathbf{w}_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + b + \sum_{i \in S} (\mathbf{x}_i \cdot \mathbf{w}_i)) = \\ step(\mathbf{w}_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + w_S \cdot \mathbf{x}_S + b) = f(\mathbf{x}_S; \mathbf{x}'_{\bar{S}}) \end{aligned} \tag{71}$$

Thus, it follows that the count of assignments for which $\mathbf{x}'_{|S|+1} = 0$ that satisfy $f'(\mathbf{x}) = 1$ precisely equals the number of assignments for which $f(\mathbf{x}_S; \mathbf{x}'_{\bar{S}}) = 1$. This is, in fact, equivalent to the *local* completion count: $C(S, f, \mathbf{x})$. Put differently, this implies that:

$$C(S, f, \mathbf{x}) = \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{72}$$

We now turn our attention to proving the second part of the Lemma. Specifically, we show that in the scenario where $f(\mathbf{x}) = 0$, the following condition is satisfied:

$$C(S, f, \mathbf{x}) = \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{73}$$

We will similarly begin by proving that, given $f(\mathbf{x}) = 0$, there exist at least $2^{|\bar{S}|}$ vectors $\mathbf{x}' \in \{0,1\}^{|S|+1}$ for which $f'(\mathbf{x}') = 0$.

First, we assume that $\mathbf{x}'_{|\bar{S}|+1} = 1$ (the assignment for feature $|\bar{S}| + 1$ is 1). Now, for all $\mathbf{x}' \in \{0,1\}^{|S|+1}$ such that $\mathbf{x}'_{|\bar{S}|+1} = 1$ it holds that:

$$\begin{aligned} \mathbf{w}' \cdot \mathbf{x}' + b' = \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \mathbf{w}'_{|\overline{S}|+1} \cdot \mathbf{x}'_{|\overline{S}|+1} + b' = \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \delta' + b' = \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} + \left(-(\sum_{i \in \overline{S}} |w_i|) - b' - 1 \right) + b' = \\ \mathbf{w}'_{\bar{S}} \cdot \mathbf{x}'_{\bar{S}} - \sum_{i \in \overline{S}} |\mathbf{w}_i| - 1 < 0 \end{aligned} \tag{74}$$

Given that there are precisely $2^{|\overline{S}|}$ assignments where $\mathbf{x}'_{|\overline{S}|+1} = 1$, it follows that there exist at least $2^{|\overline{S}|}$ assignments for which $f'(\mathbf{x}') = 0$. Due to the same reasoning as in the previously discussed case where $f'(\mathbf{x}') = 1$, it follows that the subsequent condition is met:

$$t = 2^{|\overline{S}|} + \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{75}$$

Therefore, the number of assignments, where $\mathbf{x}'_{|S|+1} = 1$, that satisfy the condition $f'(\mathbf{x}) = 0$ is as follows:

$$\sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{76}$$

Given the aforementioned reasons, we can deduce again that: $f'(\mathbf{x}'_{\bar{S}}; \mathbf{0}_{|\bar{S}|+1}) = f(\mathbf{x}_S; \mathbf{x}'_{\bar{S}})$. Consequently, the number of assignments where $\mathbf{x}'_{|S|+1} = 0$ and $f'(\mathbf{x}) = 0$ coincides with those where $f(\mathbf{x}_S; \mathbf{x}'_{\bar{S}}) = 0$. This corresponds to the *local* completion count: $C(S, f, \mathbf{x})$ in this context. In other words, it again holds that:

$$C(S, f, \mathbf{x}) = \sqrt{\frac{1}{2} \cdot C(\emptyset, f') - 2^{2|\overline{S}|}} \tag{77}$$

which concludes the reduction.

Lemma 26 *G-CC is $\#P$-Complete for MLPs.*

Proofs of membership and Hardness for Perceptrons will also clearly hold for MLPs.

Lemma 27 *G-CC is in PTIME for FBDDs.*

Proof. Similarly to the proof of the complexity of *G-CC* for Perceptrons (Lemma 24), we will assume the normalized count $C(S, f)$ which is interchangeable with $c(S, f)$. Each leaf node v of f corresponds to some path α. We denote by α_S the subset of nodes from α that correspond to the features of S. We suggest the following polynomial algorithm: We enumerate pairs of leaf nodes by iterating over all pairs in its cartesian product. In other words, given that $L := \{v_1, v_2\}$ includes all the leaf nodes of a function f, we iterate over every possible pair of leaf nodes in the set $L := \{v_1, v_2\}$. This includes the pairs $\{(v_1, v_1), (v_1, v_2), (v_2, v_1), (v_2, v_2)\}$. Given the pair (v, v'), we perform a counting procedure iff there do not exist two nodes $v_\alpha \in \alpha_S$ and $v_{\alpha'} \in \alpha'_S$ with the same input feature i and with *different* path output edges. Intuitively, this means that α and α' do not match on the subset S.

We define w.l.o.g. that v corresponds to the counting procedure over $\mathbf{x} \in \mathbb{F}$ and that v' corresponds to the counting procedure over $\mathbf{z} \in \{0, 1\}^{|\overline{S}|}$. Therefore, for each counting procedure, we add $2^{n-|\alpha|} \cdot 2^{|\overline{S}|-|\alpha'_{\bar{S}}|}$. Upon completing the iteration across all pairs (v, v'), we derive $C(S, f)$. Similarly to our previous assertions, we again note that the expression $2^{n-|\alpha|} \cdot 2^{|\overline{S}|-|\alpha'_{\bar{S}}|}$ can be computed in polynomial time, since n, the number of input features, is given in unary. Furthermore, we observe that the process of iterating through the cartesian product of leaf nodes arises from the definition of the global completion count, which remains unaffected by the order of the features.

K. Framework Extensions

Discrete and continuous input and output domains. To simplify the comprehension of our proofs, we followed common conventions (Barceló et al., 2020; Arenas et al., 2022; Wäldchen et al., 2021; Arenas et al., 2021a) and provided them over boolean input and output values. First, we observe that the proofs of duality and uniqueness presented in the initial sections are independent of any assumptions about the input or output domains, making them applicable to both discrete and continuous domains.

Next, we turn our attention to the outcomes of our complexity analysis. It is important to note that the analysis we have carried out is not limited to binary features but can also be extended to features that take on k possible values, where k represents any integer. Moreover, an additional extension can adapt our approach to incorporate continuous inputs. We will now briefly elaborate on the diverse situations in which this extension remains relevant.

Regarding MLP explainability queries, earlier research indicates that the complexity of a *satisfiability* query on an MLP extends to scenarios involving continuous inputs. Specifically, the work of (Katz et al., 2017) and (Sälzer & Lange, 2021) proves that verifying an arbitrary satisfiability query on an MLP with ReLU activations, over a continuous input domain, remains NP-complete. The *CSR* query mentioned in this work, when $S := \emptyset$ is akin to *negating* a satisfiability query, and this implies that the *CSR* query (and consequently also *G-CSR*) in MLPs remains coNP complete for the continuous case as well. We recall that the complexity of the *MSR* and *FR* queries for MLPs are Σ_2^P-Complete and Π_2^P-Complete, respectively. This complexity arises from the use of a coNP oracle, which determines whether a subset of features is sufficient, essentially addressing the *CSR* query. Given that *CSR* can also be adapted to handle continuous outputs, the logic applied to *CSR* can similarly be applied to demonstrate that both *MSR* and *FR* queries can be extended to continuous domains.

For Perceptrons, the completeness proofs remain valid in a continuous domain for *(G)-CSR*, *(G)-MSR*, and *(G)-FR* explainability queries. The continuity of inputs does not alter the membership proofs, for the same reasons that hold for MLPs. For hardness proofs, notice that all reductions that were derived from the SSP problem, can be adjusted to substitute any call to $\max\{z_i, z_j\}$ in our original proof with $\max([z_i, z_j])$.

The *FN* algorithms remain correct for the continuous scenario for Perceptrons and FBDDs, as the insight from theorem 1 persists. Thus, the algorithm recommended for decision trees remains applicable. For Perceptrons, one can directly determine the minimal and maximal viable values by fixing $S \setminus i$ and allowing i to take any possible value. In the context of MLPs, solving the problem remains within polynomial time complexity for discrete inputs, yet transitions to be NP-Complete when dealing with continuous cases. Moreover, it is important to acknowledge that the proofs for the *CC* query also hold only for the discrete case, as its inherent counting nature renders it undefined for the continuous version too.

Finally, the proofs that apply to tree classifiers for queries such as *(G)-CSR*, *(G)-MSR*, *(G)-FN*, and *(G)-FR* are equally valid for continuous inputs. This extension to continuous inputs for the local forms mentioned was demonstrated by previous work (Huang et al., 2021). Given that the global algorithms for decision trees resemble their local counterparts, with the distinction of enumerating pairs of leaf nodes instead of single leaf nodes, the correctness of these algorithms persists in the continuous domain. Consequently, their complexity remains polynomial.

Relaxations, probabilistic classification, and regression. Other possible extensions of our framework might involve alternative, more flexible definitions of our explanation forms, as well as adaptations to different contexts like probabilistic classification or regression.

Possible relaxations of our definitions could incorporate *probabilistic* notions of sufficiency (Wäldchen et al., 2021; Izza et al., 2021; Arenas et al., 2022; Wang et al., 2021a), restricting them to bounded ϵ-ball domains (Malfa et al., 2021; Wu et al., 2024b; Izza et al., 2024; Huang & Marques-Silva, 2023), or to meaningful distributions (Yu et al., 2023; Gorji & Rubin, 2022). Additionally, our definitions could be adapted from the simplified binary classification to regression or probabilistic classification contexts. For instance, in the case of a neural network regression model $f : \mathbb{F} \to \mathbb{R}$, a *sufficient reason* might be defined as a subset $S \subseteq \{1, \ldots, n\}$ of input features such that:

$$\forall \mathbf{z} \in \mathbb{F} \quad ||f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) - f(\mathbf{x})||_p \leq \delta \tag{78}$$

for some $0 \leq \delta \leq 1$ and some ℓ_p-norm. Other concepts explored in our work, such as contrastive reasons, the global versions of our explanation definitions, and the related definitions of necessity and redundancy, can be similarly adapted to this framework.

L. Terminology and Relationship to Other Explanation Forms

In this section, we will outline several definitions from the literature that are relevant to those discussed here.

Sufficient reasons and abductive explanations. A sufficient reason is also commonly referred to as an *abductive explanation* (Ignatiev et al., 2019a) (abbreviated as *AXP*) or a *prime implicant* (Shih et al., 2018) (abbreviated as *PI-explanation*). Importantly, our definition of a sufficient reason does not automatically imply subset minimality, which is sometimes the case in other works. For instance, (Marques-Silva & Ignatiev, 2022) describe an AXP as a subset-minimal sufficient subset of features, distinguishing it from a *weak*-AXP, where subset minimality does not necessarily hold. We follow the conventions used by (Barceló et al., 2020), and define a sufficient reason as any sufficient subset of features, and we specify that a subset S is a *subset-minimal* sufficient reason only when it is explicitly so.

Absloute vs. global sufficient reasons. Here, we highlight the distinctions between *global* sufficient reasons, as discussed throughout this work, and the notion of *absolute* sufficient reasons outlined in (Ignatiev et al., 2019b), in which the authors refer to these as *absolute explanations*, while (Marques-Silva, 2023) define them as global abductive explanations. For a given function f and a prediction class $c \in \{0, 1\}$, an absolute sufficient reason with respect to $\langle f, c \rangle$ is a *partial assignment* $\mathbf{x}_S \in \{0, 1\}^{|S|}$ to the features in S, such that:

$$\forall \mathbf{z} \in \mathbb{F} \ [\ f(\mathbf{x}_S; \mathbf{z}_{\bar{S}}) = c\ ] \tag{79}$$

To determine whether a specific instance qualifies as an *absolute* sufficient reason, we would need to evaluate a *partial assignment* to an input $\mathbf{x}_S$. However, this work centers on comparing local explanations, which apply to a particular instance

$\mathbf{x}$, to their corresponding global explanations, which apply to *any* possible instance $\mathbf{x}$. In this context, we concentrate on either the local scenario, which involves subsets of features that are sufficient for determining a specific prediction for $\mathbf{x}$, or the global scenario, where subsets of features are sufficient to determine the prediction for *any* $\mathbf{x}$. As demonstrated by proposition 5, this definition of global sufficient reasons also connects to the concepts of necessity and redundancy of features. Specifically, the provably unique subset-minimal global sufficient reason for a function f categorizes features into those that are necessary for a specific $\langle f, \mathbf{x}\rangle$ and those that are globally redundant.

Necessity, redundancy, and bias detection. We follow the notions of feature necessity and redundancy as discussed in (Huang et al., 2023), where the focus is primarily on the *local* versions of these explanations. In contrast, our analysis extends to both local variants, which apply to a specific instance $\mathbf{x}$, and global variants, which apply to any instance $\mathbf{x}$.

These notions of necessity correspond to those discussed in (Darwiche, 2023; Darwiche & Ji, 2022; Watson et al., 2021), while the ideas of redundancy are related to fairness and bias, as explored in other studies (Darwiche & Hirth, 2020; Arenas et al., 2021a; Ignatiev et al., 2020a). There, it is often considered that there exists a set $P \subseteq \{1, \ldots, n\}$ of protected features that should not influence the prediction. Notably, (Ignatiev et al., 2020a) apply the criterion of fairness through unawareness (FTU), which involves ensuring that all features in P are redundant, whether locally or concerning a specific prediction class. Similarly, (Darwiche & Hirth, 2020) differentiate between *local* biases, termed *prediction* bias, and *global* biases, referred to as *classifier* bias. More specifically, a *prediction* $\langle f, \mathbf{x}\rangle$ is biased iff:

$$\exists \mathbf{z} \in \mathbb{F} \ \left[f(\mathbf{x}_{\bar{P}}; \mathbf{z}_P) \neq f(\mathbf{x}) \right] \tag{80}$$

which is equivalent to P being *locally* contrastive with respect to $\langle f, \mathbf{x}\rangle$. In the *global* context, a classifier f is deemed biased iff there is at least one input $\mathbf{x}$ where the prediction $\langle f, \mathbf{x}\rangle$ is biased. Equivalently, a classifier f is *unbiased* if and only if all protected features in P are *globally redundant*. A similar notation of bias detection is discussed in (Arenas et al., 2021a). Moreover, (Audemard et al., 2021) explore a modified version of this concept, focusing on irrelevancy (redundancy) concerning a particular class.

Counting completions and δ-relevant sets. In this work, we investigate the computational complexity of the Count-Completion (*CC*) query, a form of explanation also examined in (Barceló et al., 2020). However, their study concentrated on the local variant of this query, whereas we address both the local and global variants. The *CC* query closely (but not exactly) aligns with a δ-relevant set (Wäldchen et al., 2021; Izza et al., 2021). When obtaining a δ-relevant set, the focus is not on calculating the completion count itself, but rather on determining whether the completion count surpasses a specified threshold δ.