Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning
Main Track

Explaining Decisions in ML Models: a Parameterized Complexity Analysis

Sebastian Ordyniak[1] , **Giacomo Paesani**[2] , **Mateusz Rychlicki**[1] , **Stefan Szeider**[3]
[1]University of Leeds, Leeds, UK
[2]Sapienza University of Rome, Rome, Italy
[3]TU Wien, Vienna, Austria
{sordyniak,giacomopaesani,mkrychlicki}@gmail.com, sz@ac.tuwien.ac.at

Abstract

This paper presents a comprehensive theoretical investigation into the parameterized complexity of explanation problems in various machine learning (ML) models. Contrary to the prevalent black-box perception, our study focuses on models with transparent internal mechanisms. We address two principal types of explanation problems: abductive and contrastive, both in their local and global variants. Our analysis encompasses diverse ML models, including Decision Trees, Decision Sets, Decision Lists, Ordered Binary Decision Diagrams, Random Forests, and Boolean Circuits, and ensembles thereof, each offering unique explanatory challenges. This research fills a significant gap in explainable AI (XAI) by providing a foundational understanding of the complexities of generating explanations for these models. This work provides insights vital for further research in the domain of XAI, contributing to the broader discourse on the necessity of transparency and accountability in AI systems.

1 Introduction

As machine learning (ML) models increasingly permeate essential domains, understanding their decision-making mechanisms has become central. This paper delves into the field of explainable AI (XAI) by examining the parameterized complexity of explanation problems in various ML models. We focus on models with accessible internal mechanisms, shifting away from the traditional black-box paradigm. Our motivation is rooted in establishing a comprehensive theoretical framework that illuminates the complexity of generating explanations for these models, a task becoming increasingly relevant in light of recent regulatory guidelines that emphasize the importance of transparent and explainable AI (Commission 2020; OECD 2023).

The need for transparency and accountability in automated decision-making drives the imperative for explainability in AI systems, especially in high-risk sectors. ML models, while powerful, must be demystified to gain trust and comply with ethical and regulatory standards. Formal explanations serve this purpose, providing a structured means to interpret model decisions (Marques-Silva 2023; Guidotti et al. 2019; Carvalho, Pereira, and Cardoso 2019).

Our exploration focuses on two types of explanation problems, abductive and contrastive, in local and global contexts (Marques-Silva 2023). *Abductive explanations* (Ignatiev, Narodytska, and Marques-Silva 2019), corresponding to prime-implicant explanations (Shih, Choi, and Darwiche 2018) and sufficient reason explanations (Darwiche and Ji 2022), clarify specific decision-making instances, while *contrastive explanations* (Miller 2019; Ignatiev et al. 2020), corresponding to necessary reason explanations (Darwiche and Ji 2022), make explicit the reasons behind the non-selection of alternatives. The study of contrastive explanations goes back to the Lipton's work in 1990. Conversely, *global explanations* (Ribeiro, Singh, and Guestrin 2016; Ignatiev, Narodytska, and Marques-Silva 2019) aim to unravel models' decision patterns across various inputs. This bifurcated approach enables a comprehensive understanding of model behavior, aligning with the recent emphasis on interpretable ML (Lisboa et al. 2023).

In contrast to a recent study by Ordyniak, Paesani, and Szeider (2023), who consider the parameterized complexity of finding explanations based on samples classified by a black-box ML model, we focus on the setting where the model together with its inner workings is available as an input for computing explanations. This perspective, initiated by Barceló et al. (2020), is particularly appealing, as it lets us quantify the explainability of various model types based on the computational complexity of the corresponding explanation problems.

Challenging the notion of inherent opacity in ML models, our study includes *Decision Trees* (DTs), *Decision Sets* (DSs), *Decision Lists* (DLs), and *Ordered Binary Decision Diagrams* (OBDDs). Whereas DTs, DSs, and DLs are classical ML models, OBDDs can be used to represent the decision, functions of naive Bayes classifiers (Chan and Darwiche 2003). We also consider *ensembles* of all the above ML models; where an ensemble classifies an example by taking the majority classification over its elements. For instance, *Random Forests* (RFs) are ensembles of DTs.

Each model presents distinct features affecting explanation generation. For example, the transparent structure of DTs and RFs facilitates rule extraction, as opposed to the complex architectures of *Neural Networks* (NNs) (Ribeiro, Singh, and Guestrin 2016; Lipton 2018).

Contribution This paper fills a crucial gap in XAI research by analyzing the complexity of generating explanations across different models. Prior research has often centred on practical explainability approaches, but a theoretical under-

standing still needs to be developed (Holzinger et al. 2020; Molnar 2023). Our study is aligned with the increasing call for theoretical rigor in AI (Commission 2019). By dissecting the parameterized complexity of these explanation problems, we lay the groundwork for future research and algorithm development, ultimately contributing to more efficient explanation methods in AI.

Since most of the considered explanation problems are NP-hard, we use the paradigm of fixed-parameter tractability (FPT), which involves identifying specific parameters of the problem (e.g., explanation size, number of terms/rules, size/height of a DT, width of a BDD) and proving that the problem is fixed-parameter tractable concerning these parameters. By focusing on these parameters, the complexity of the problem is confined, making it more manageable and often solvable in uniform polynomial time for fixed values of the parameters. A significant part of our positive results are based on reducing various model types to *Boolean circuits (*BC*s)*. This reduction is crucial for the uniform treatment of several model types as it allows the application of known algorithmic results and techniques from the Boolean circuits domain to the studied models. It simplifies the problems and brings them into a well-understood theoretical framework. For ensembles, we consider Boolean circuits with majority gates. In turn, we obtain the fixed-parameter tractability of problems on Boolean circuits via results on Monadic Second Order (MSO). We use extended MSO (Bergougnoux, Dreier, and Jaffke 2023) to handle majority gates, which allows us to obtain efficient algorithmic solutions, particularly useful for handling complex structures.

Overall, the approach in the manuscript is characterized by a mix of theoretical computer science techniques, including parameterization, reduction to well-known problems, and the development of specialized algorithms that exploit the structural properties of the models under consideration. This combination enables the manuscript to effectively address the challenge of finding tractable solutions to explanation problems in various machine learning models.

For some of the problems, we develop entirely new customized algorithms. We complement the algorithmic results with hardness results to get a complete picture of the tractability landscape for all possible combinations of the considered parameters (an overview of our results are provided in Tables 2, 3, 4).

In summary, our research marks a significant advancement in the theoretical understanding of explainability in AI. By offering a detailed complexity analysis for various ML models, this work enriches academic discourse and responds to the growing practical and regulatory demand for transparent, interpretable, and trustworthy AI systems.

A full version of the paper can be found on ArXiv (Ordyniak et al. 2024).

2 Preliminaries

Parameterized Complexity. A problem with input size n and parameter k is *fixed-parameter tractable (fp-tractable)* if it can be solved in time $f(k)n^c$ for a constant c independent of k, and a computable function f; the problem is *xp-tractable* if it can be solved in time $n^{f(k)}$ (Downey and Fellows 2013). FPT and XP are the classes of fp-tractable and xp-tractable decision problems, respectively. There is a hierarchy of parameterized complexity classes that represent various levels of *in*tractability: $\mathsf{P} \subseteq \mathsf{FPT} \subseteq \mathsf{W}[1] \subseteq \mathsf{W}[2] \subseteq \cdots \subseteq \mathsf{XP} \cap \mathsf{paraNP} \subseteq \mathsf{paraNP}$. All inclusions are believed to be proper. If a problem is $\mathsf{W}[i]$-hard under fpt-reductions ($\mathsf{W}[i]$-h, for short) then it is unlikely to be in FPT. The class of $\mathsf{co\text{-}C}$ denotes the complexity class containing all problems from C with yes-instances and no-instances swapped.

Examples and Models Let F be a set of binary features. An *example* $e : F \to \{0, 1\}$ over F is a $\{0, 1\}$-assignment of the features in F. An example is a *partial example (assignment)* over F if it is an example over some subset F' of F. We denote by $E(F)$ the set of all possible examples over F. A *(binary classification) model* $M : E(F) \to \{0, 1\}$ is a specific representation of a Boolean function over $E(F)$. We denote by $F(M)$ the set of features considered by M, i.e., $F(M) = F$. We say that an example e is a 0-example or negative example (1-example or positive example) w.r.t. the model M if $M(e) = 0$ ($M(e) = 1$). For convenience, we restrict our setting to the classification into two classes. We note however that all our hardness results easily carry over to the classification into any (in)finite set of classes. The same applies to our algorithmic results for non-ensemble models since one can easily reduce to the case with two classes by renaming the class of interest for the particular explanation problem to 1 and all other classes to 0. We leave it open whether the same holds for our algorithmic results for ensemble models.

Decision Trees. A *decision tree* (DT) $\mathcal{T}$ is a pair (T, λ) such that T is a rooted binary tree and $\lambda : V(T) \to F \cup \{0, 1\}$ is a function that assigns a feature in F to every inner node of T and either 0 or 1 to every leaf node of T. Every inner node of T has exactly 2 children, one left child (or 0-child) and one right-child (or 1-child). The classification function $\mathcal{T} : E(F) \to \{0, 1\}$ of a DT is defined as follows for an example $e \in E(F)$. Starting at the root of T one does the following at every inner node t of T. If $e(\lambda(t)) = 0$ one continues with the 0-child of t and if $e(\lambda(t)) = 1$ one continues with the 1-child of t until one eventually ends up at a leaf node l at which e is classified as $\lambda(l)$. For every node t of T, we denote by $\alpha^t_{\mathcal{T}}$ the partial assignment of F defined by the path from the root of T to t in T, i.e., for a feature f, we set $\alpha^t_{\mathcal{T}}(f)$ to 0 (1) if and only if the path from the root of T to t contains an inner node t' with $\lambda(t') = f$ together with its 0-child (1-child). We denote by $L(\mathcal{T})$ the set of leaves of T and we set $L_b(\mathcal{T}) = \{\, l \in L(\mathcal{T}) \mid \lambda(l) = b \,\}$ for every $b \in \{0, 1\}$. Moreover, we denote by $\|\mathcal{T}\|$ ($h(\mathcal{T})$) the size (height) of a DT, which is equal to the number of leaves of T (the length of a longest root-to-leaf path in T). Finally, we let $\mathsf{MNL}(\mathcal{T}) = \min\{|L_0|, |L_1|\}$.

Decision Sets. A *term* t over F is a set of *literals* with each literal being of the form $(f = z)$ where $f \in F$ and $z \in \{0, 1\}$. A *rule* r is a pair (t, c) where t is a term and $c \in \{0, 1\}$. We say that a rule (t, c) is a *c-rule*. We say that a term t (or rule (t, c)) *applies to (or agrees with)* an example

e if $e(f) = z$ for every element $(f = z)$ of t. Note that the empty rule applies to any example.

A *decision set* (DS) S is a pair (T, b), where T is a set of terms and $b \in \{0, 1\}$ is the classification of the default rule (or the default classification). We denote by $\|S\|$ the size of S which is equal to $(\sum_{t \in T} |t|) + 1$; the $+1$ is for the default rule. The classification function $S : E(F) \to \{0, 1\}$ of a DS $S = (T, b)$ is defined by setting $S(e) = b$ for every example $e \in E(F)$ such that no term in T applies to e and otherwise we set $S(e) = 1 - b$.

Decision Lists. A *decision list* (DL) L is a non-empty sequence of rules $(r_1 = (t_1, c_1), \ldots, r_\ell = (t_\ell, c_\ell))$, for some $\ell \geq 0$. The size of a DL L, denoted by $\|L\|$, is equal to $\sum_{i=1}^{\ell}(|t_i| + 1)$. The classification function $L : E(F) \to \{0, 1\}$ of a DL L is defined by setting $L(e) = b$ if the first rule in L that applies to e is a b-rule. To ensure that every example obtains some classification, we assume that the term of the last rule is empty and therefore applies to all examples.

Binary Decision Diagrams. A *binary decision diagram* (BDD) B is a pair (D, ρ) where D is a directed acyclic graph with three special vertices s, t_0, t_1 such that:

- s is a source vertex that can (but does not have to) be equal to t_0 or t_1,
- t_0 and t_1 are the only sink vertices of D,
- every non-sink vertex has exactly two outgoing neighbors, which we call the 0-neighbor and the 1-neighbor, and
- $\rho : V(D) \setminus \{t_0, t_1\} \to F$ is a function that associates with every non-sink node of D a feature in F.

For an example $e \in E$, we denote by $P_B(e)$ (or $P(e)$ if B is clear from the context), the unique path from s to either t_0 or t_1 followed by e in B. That is starting at s and ending at either t_0 or t_1, $P(e)$ is iteratively defined as follows. Initially, we set $P(e) = (s)$, moreover, if $P(e)$ ends in a vertex v other than t_0 or t_1, then we extend $P(e)$ by the $e(\rho(v))$-neighbor of v in D. Let B be a BDD and $e \in E(F)$ be an example. The classification function $B : E(F) \to \{0, 1\}$ of B is given by setting $B(e) = b$ if $P_B(e)$ ends in t_b. We denote by $\|B\|$ the size of B, which is equal to $|V(D)|$. We say that B is an OBDD if every path in B contains features in the same order. Moreover, B is a *complete* OBDD if every maximal path contains the same set of features. It is known that every OBDD can be transformed in polynomial-time into an equivalent complete OBDD (Mengel and Slivovsky 2021, Observation 1). All OBDDs considered in the paper are complete.

Ensembles. An $\mathcal{M}$-*ensemble*, also denoted by $\mathcal{M}_{\text{MAJ}}$, $\mathcal{E}$ is a set of models of type $\mathcal{M}$, where $\mathcal{M} \in \{\text{DT}, \text{DS}, \text{DL}, \text{OBDD}\}$. We say that $\mathcal{E}$ classifies an example $e \in E(F)$ as b if so do the majority of models in $\mathcal{E}$, i.e., if there are at least $\lfloor |\mathcal{E}|/2 \rfloor + 1$ models in $\mathcal{E}$ that classify e as b. We denote by $\|\mathcal{E}\|$ the size of $\mathcal{E}$, which is equal to $\sum_{M \in \mathcal{E}} \|M\|$. We additionally consider an *ordered* OBDD-*ensemble*, denoted by $\text{OBDD}^{<}_{\text{MAJ}}$, where all OBDDs in the ensemble respect the same ordering of the features.

r_1 : IF	$(x = 1 \wedge y = 1)$	THEN 0
r_2 : ELSE IF	$(x = 0 \wedge z = 0)$	THEN 1
r_3 : ELSE IF	$(y = 0 \wedge z = 1)$	THEN 0
r_4 : ELSE		THEN 1

Figure 1: Let L be the DL given in the figure and let e be the example given by $e(x) = 0$, $e(y) = 0$ and $e(z) = 1$. Note that $L(e) = 0$. It is easy to verify that $\{y, z\}$ is the only local abductive explanation for e in L of size at most 2. Moreover, both $\{y\}$ and $\{z\}$ are minimal local contrastive explanations for e in L. Let $\tau_1 = \{x \mapsto 1, y \mapsto 1\}$ and $\tau_2 = \{x \mapsto 0, z \mapsto 0\}$ be a partial assignments. Note that τ_1 and τ_2 are minimal global abductive and global contrastive explanations for class 0 w.r.t. L, respectively.

3 Considered Problems and Parameters

We consider the following types of explanations (see Marques-Silva's survey (2023)). Let M be a model, e an example over $F(M)$, and let $c \in \{0, 1\}$ be a classification (class). We consider the following types of explanations for which an example is illustrated in Figure 1.

- A *(local) abductive explanation (*LAXP*) for e w.r.t. M* is a subset $A \subseteq F(M)$ of features such that $M(e) = M(e')$ for every example e' that agrees with e on A.
- A *(local) contrastive explanation (*LCXP*) for e w.r.t. M* is a set A of features such that there is an example e' such that $M(e') \neq M(e)$ and e' differ from e only on the features in A.
- A *global abductive explanation (*GAXP*) for c w.r.t. M* is a partial example $\tau : F \to \{0, 1\}$, where $F \subseteq F(M)$, such that $M(e) = c$ for every example e that agrees with τ.
- A *global contrastive explanation (*GCXP*) for c w.r.t. M* is a partial example $\tau : F \to \{0, 1\}$, where $F \subseteq F(M)$, such that $M(e) \neq c$ for every example that agrees with τ.

For each of the above explanation types, each of the considered model types $\mathcal{M}$, and depending on whether or not one wants to find a subset minimal or cardinality-wise minimum explanation, one can now define the corresponding computational problem. For instance:

$\mathcal{M}$-SUBSET-MINIMAL LOCAL ABDUCTIVE EXPLANATION (LAXP$_{\subseteq}$)
INSTANCE: A model $M \in \mathcal{M}$ and an example e.
QUESTION: Find a subset minimal local abductive explanation for e w.r.t. M.

$\mathcal{M}$-CARDINALITY-MINIMAL LOCAL ABDUCTIVE EXPLANATION (LAXP$_{||}$)
INSTANCE: A model $M \in \mathcal{M}$, an example e, and an integer k.
QUESTION: Is there a local explanation for e w.r.t. M of size at most k?

The problems $\mathcal{M}$-$X_{\subseteq}$ and $\mathcal{M}$-$X_{||}$ for $X \in \{$GAXP, LCXP, GCXP$\}$ are defined analogously.

Finally, for these problems, we will consider natural parameters listed in Table 1; not all parameters apply to all

considered problems. We denote a problem X parameterized by parameters p, q, r by $X(p+q+r)$.

4 Overview of Results

As we consider several problems, each with several variants and parameters, there are hundreds of combinations to consider. We therefore provide a condensed summary of our results in Tables 2, 3, 4.

The first column in each table indicates whether a result applies to the cardinality-minimal or subset-minimal variant of the explanation problem (i.e., to $X_{\subseteq}$ or $X_{|\cdot|}$, respectively). The next 4 columns in Tables 2, 3, 4 indicate the parameterization, the parameters are explained in Table 1. A "p" indicates that this parameter is part of the parameterization, a "$-$" indicates that it isn't. A "c" means the parameter is set to a constant, "1" means the constant is 1.

By default, each row in the tables applies to all four problems LAXP, GAXP, GCXP, and LCXP. However, if a result only applies to LCXP, it is stated in parenthesis. So, for instance, the first row of Table 2 indicates that DT-LAXP$_{\subseteq}$, DT-GAXP$_{\subseteq}$, DT-GCXP$_{\subseteq}$, and DT-LCXP$_{\subseteq}$, where the ensemble consists of a single DT, can be solved in polynomial time.

The penultimate row of Table 2 indicates that DT$_{\text{MAJ}}$-LAXP$_{|\cdot|}$, DT$_{\text{MAJ}}$-GAXP$_{|\cdot|}$ and DT$_{\text{MAJ}}$-GCXP$_{|\cdot|}$ are co-NP-hard even if *mnl_size* + *size_elem* + *xp_size* is constant, and DT$_{\text{MAJ}}$-LCXP$_{|\cdot|}$ is W[1]-hard parameterized by *xp_size* even if *mnl_size* + *size_elem* is constant. Finally, the $\star$ indicates a minor distinction in the complexity between DT-LAXP$_{|\cdot|}$ and the two problems DT-GAXP$_{|\cdot|}$ and DT-GCXP$_{|\cdot|}$. That is, if the cell contains NP-h* or pNP-h*, then DT-LAXP$_{|\cdot|}$ is NP-hard or pNP-hard, respectively, and neither DT-GAXP$_{|\cdot|}$ nor DT-GCXP$_{|\cdot|}$ are in P unless FPT = W[1].

We only state in the tables those results that are not implied by others. Tractability results propagate in the following list from left to right, and hardness results propagate from right to left.

$\mid\!\cdot\!\mid$-minimality	$\Rightarrow$	$\subseteq$-minimality
set A of parameters	$\Rightarrow$	set $B \supseteq A$ of parameters
ensemble of models	$\Rightarrow$	single model
unordered OBDD ensemble	$\Rightarrow$	ordered OBDD ensemble

For instance, the tractability of $X_{|\cdot|}$ implies the tractability of $X_{\subseteq}$, and the hardness of $X_{\subseteq}$ implies the hardness of $X_{|\cdot|}$.

parameter	definition
ens_size	number of elements of the ensemble
mnl_size	largest number of MNL over all ensemble elem.
terms_elem	largest number of terms per ensemble elem.
term_size	size of a largest term over all ensemble elem.
width_elem	largest width over all ensemble elements
size_elem	size of largest ensemble element
xp_size	size of the explanation

Table 1: Main parameters considered. Note that some parameters (such as *width_elem*) only apply to specific model types.

minimality	*ens_size*	*mnl_size*	*size_elem*	*xp_size*	complexity	result
$\subseteq$	1	$-$	$-$	$-$	P	Thm 5
$\mid\!\cdot\!\mid$	1	$-$	$-$	$-$	NP-h*(P)	Thms 5, 20, 21
$\mid\!\cdot\!\mid$	1	$-$	$-$	p	W[1]-h(P)	Thms 5, 20, 21
$\mid\!\cdot\!\mid$	1	$-$	$-$	p	XP (P)	Thms 5, 6
$\subseteq$	p	$-$	$-$	$-$	co-W[1]-h(W[1]-h)	Thm 23
$\subseteq$	p	$-$	$-$	$-$	XP	Thm 7
$\mid\!\cdot\!\mid$	p	$-$	$-$	$-$	pNP-h*(XP)	Thms 7, 20, 21
$\mid\!\cdot\!\mid$	p	p	$-$	$-$	FPT	Thm 3
$\mid\!\cdot\!\mid$	p	$-$	p	$-$	FPT	Thm 3
$\mid\!\cdot\!\mid$	p	$-$	$-$	c(p)	co-W[1]-h(W[1]-h)	Thm 23
$\subseteq$	$-$	c	c	$-$	co-NP-h(NP-h)	Thm 24
$\mid\!\cdot\!\mid$	$-$	c	c	c(p)	co-NP-h(W[1]-h)	Thm 24
$\mid\!\cdot\!\mid$	$-$	$-$	$-$	p	co-pNP-h(XP)	Thms 1, 24

Table 2: Explanation complexity when the model is a DT or an ensemble of DTs. See Section 4 for how to read the table.

minimality	*ens_size*	*terms_elem*	*term_size*	*xp_size*	complexity	result
$\subseteq$	1	$-$	c	$-$	co-NP-h(NP-h)	Thm 25
$\mid\!\cdot\!\mid$	1	$-$	$-$	p	co-pNP-h(W[1]-h)	Thms 25, 26
$\mid\!\cdot\!\mid$	c	$-$	p	p	co-pNP-h(FPT)	Thms 11, 25
$\mid\!\cdot\!\mid$	p	p	$-$	$-$	FPT	Cor 8
$\mid\!\cdot\!\mid$	p	$-$	c	p	co-pNP-h(W[1]-h)	Thms 25, 27
$\subseteq$	$-$	c	c	$-$	co-NP-h(NP-h)	Thm 28
$\mid\!\cdot\!\mid$	$-$	c	c	c(p)	co-NP-h(W[1]-h)	Thm 28
$\mid\!\cdot\!\mid$	$-$	$-$	$-$	p	co-pNP-h(XP)	Thms 1, 25

Table 3: Explanation complexity when the model is a DS, a DL, or an ensemble thereof. See Section 4 for how to read the table.

5 Algorithmic Results

In this section, we will present our algorithmic results. We start with some general observations that are independent of a particular model type.

Theorem 1. *Let $\mathcal{M}$ be any model type such that $M(e)$ can be computed in polynomial-time for $M \in \mathcal{M}$. $\mathcal{M}$-LCXP$_{|\cdot|}$ parameterized by xp_size is in* XP.

Proof. Let (M, e, k) be the given instance of $\mathcal{M}$-LCXP$_{|\cdot|}$ and suppose that $A \subseteq F(M)$ is a cardinality-wise minimal local contrastive explanation for e w.r.t. M. Because A is cardinality-wise minimal, it holds the example e_A obtained from e by setting $e_A(f) = 1 - e(f)$ for every $f \in A$ and $e_A(f) = e(f)$ otherwise, is classified differently from e, i.e., $M(e) \neq M(e_A)$. Therefore, a set $A \subseteq F(M)$ is a cardinality-wise minimal local contrastive explanation for e w.r.t. M if and only if $M(e) \neq M(e_A)$ and there is no cardinality-wise smaller set A' for which this is the case. This now allows us to obtain an XP algorithm for $\mathcal{M}$-LCXP$_{|\cdot|}$ as follows. We first enumerate all possible subsets $A \subseteq F(M)$ of size at most k in time $\mathcal{O}(|F(M)|^k)$

minimality	ordered/unordered	*ens_size*	*width_elem*	*size_elem*	*xp_size*	complexity	result
$\subseteq$	u	1	–	–	–	P	Thm 14
$=$	o	1	–	–	–	NP-h(P)	Thms 14, 30
$=$	o	1	–	–	p	W[2]-h(P)	Thms 14, 30
$\subseteq$	u	c	c	–	–	co-NP-h(NP-h)	Thm 29
$=$	u	c	c	–	c(p)	co-NP-h(W[1]-h)	Thm 29
$\subseteq$	o	p	–	–	–	co-W[1]-h(W[1]-h)	Thm 32
$\subseteq$	o	p	–	–	–	XP	Thm 15
$=$	o	p	–	–	–	pNP-h(XP)	Thms 15, 30
$=$	o	p	p	–	–	FPT	Cor 12
$=$	u	p	–	p	–	FPT	Cor 13
$=$	o	p	–	–	c(p)	co-W[1]-h(W[1]-h)	Thm 32
$\subseteq$	o	–	c	c	–	co-NP-h(NP-h)	Thm 31
$=$	o	–	c	c	c(p)	co-NP-h(W[1]-h)	Thm 31
$=$	u	–	–	–	p	co-pNP-h(XP)	Thms 1, 29

Table 4: Explanation complexity when the model is an OBDD or an ensemble thereof. For an ensemble, column "ordered/unordered" indicates whether all the OBDDs in the ensemble have the same variable-order. See Section 4 for how to read the table.

and for each such subset A we test in polynomial-time if $M(e_A) \neq M(e)$. If so, we output that (M, e, k) is a yes-instance and if this is not the case for any of the enumerated subsets, we output correctly that (M, e, k) is a no-instance. □

The remainder of the section is organized as follows. First in Section 5.1, we provide a very general result about Boolean circuits, which will allow us to show a variety of algorithmic results for our models. We then provide our algorithms for the considered models in Subsections 5.2 to 5.4

5.1 A Meta-Theorem for Boolean Circuits

Here, we present our algorithmic result for Boolean circuits that are allowed to employ majority circuits. In particular, we will show that all considered explanation problems are fixed-parameter tractable parameterized by the so-called rankwidth of the Boolean circuit as long as the Boolean circuit uses only a constant number of majority gates; see, e.g., (Oum and Seymour 2006) for a definition of rankwidth. Since our considered models can be naturally translated into Boolean circuits, which require majority gates in the case of ensembles, we will obtain a rather large number of algorithmic consequences from this result by providing suitable reductions of our models to Boolean circuits in the following subsections.

To show our algorithmic result for Boolean circuits given below, we make use of an only recently developed meta-theorem (Bergougnoux, Dreier, and Jaffke 2023, Theorem 1.2) involving an extension of Monadic second order logic that allows us to easily model majority gates of Boolean circuits.

Theorem 2. *c-BC-LAXP$_{||}$, c-BC-GAXP$_{||}$, c-BC-LCXP$_{||}$, c-BC-GCXP$_{||}$ are fixed-parameter tractable parameterized by the rankwidth of the circuit.*

5.2 DTs and their Ensembles

Here, we present our algorithms for DTs and their ensembles. With the help of our meta-theorem (Theorem 2) together with natural translations of DTs and DT$_{\text{MAJ}}$s into BCs and 1-BCs, respectively, we obtain the following two theorems, showing that all problems are fixed-parameter tractable parameterized by *ens_size* plus *mnl_size*.

Theorem 3. *Let $\mathcal{P} \in \{$LAXP, LCXP, GAXP, GCXP$\}$. DT$_{\text{MAJ}}$-$\mathcal{P}_{||}$(ens_size + mnl_size) and therefore also DT$_{\text{MAJ}}$-$\mathcal{P}_{||}$(ens_size + size_elem) is* FPT.

The following auxiliary lemma provides polynomial-time algorithms for testing whether a given subset of features A is a local abductive, global abductive, or global contrastive explanation.

Lemma 4. *Let $\mathcal{T}$ be a* DT*, let e be an example and let c be a class. There are polynomial-time algorithms for the following problems:*

(1) Decide whether a given subset $A \subseteq F(\mathcal{T})$ of features is a local abductive explanation for e w.r.t. $\mathcal{T}$.

(2) Decide whether a given partial example e' is a global abductive/contrastive explanation for c w.r.t. $\mathcal{T}$.

Proof Sketch. Let $\mathcal{T}$ be a DT, let e be an example and let c be a class. Note that we assume here that $\mathcal{T}$ does not have any contradictory path.

We start by showing (1). A subset $A \subseteq F(\mathcal{T})$ of features is a local abductive explanation for e w.r.t. $\mathcal{T}$ if and only if the DT $\mathcal{T}_{|e_{|A}}$ does only contain $\mathcal{T}(e)$-leaves, which can clearly be decided in polynomial-time. Here, $e_{|A}$ is the partial example equal to the restriction of e to A. Moreover, $\mathcal{T}_{|e'}$ for a partial example e' is the DT obtained from $\mathcal{T}$ after removing every $1 - e'(f)$-child from every node t of $\mathcal{T}$ assigned to a feature f for which e' is defined. The proof for (2) is similar. □

Using dedicated algorithms for the inclusion-wise minimal variants of LAXP, GAXP, GCXP and using the polynomial-time algorithm for the cardinality-wise minimal version of LCXP given in (Barceló et al. 2020, Lemma 14), we obtain the following result.

Theorem 5. *Let $\mathcal{P} \in \{$LAXP, LCXP, GAXP, GCXP$\}$. DT-$\mathcal{P}_{\subseteq}$ and DT-LCXP$_{||}$ can be solved in polynomial-time.*

Proof Sketch. Note that the statement of the theorem for DT-LCXP$_{\subseteq}$ follows immediately from (Barceló et al. 2020, Lemma 14). Therefore, it suffices to show the statement of the theorem for the remaining 3 problems.

The polynomial-time algorithm for an instance $(\mathcal{T}, c)$ of DT-GAXP$_{\subseteq}$ works as follows. Let l be a c-leaf of $\mathcal{T}$; if no such c-leaf exists, then we can correctly output that there is no global abductive explanation for c w.r.t. $\mathcal{T}$. Then, $\alpha^l_{\mathcal{T}}$ is a global abductive explanation for c w.r.t. $\mathcal{T}$. To obtain an inclusion-wise minimal solution, we do the following. Let $F = F(\alpha^l_{\mathcal{T}})$ be the set of features on which $\alpha^l_{\mathcal{T}}$ is defined.

We now test for every feature $f \in F$ whether the restriction $\alpha^l_{\mathcal{T}}[F \setminus \{f\}]$ of $\alpha^l_{\mathcal{T}}$ to $F \setminus \{f\}$ is a global abductive explanation for c w.r.t. $\mathcal{T}$. This can clearly be achieved in polynomial-time with the help of Lemma 4. If this is true for any feature $f \in F$, then we repeat the process for $\alpha^l_{\mathcal{T}}[F \setminus \{f\}]$, otherwise we output $\alpha^l_{\mathcal{T}}$. Very similar algorithms now also work for DT-GCXP$_{\subseteq}$ and DT-LAXP$_{\subseteq}$. □

The following theorem uses an exhaustive enumeration of all possible explanations together with Lemma 4 to check whether a set of features or a partial example is an explanation.

Theorem 6. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. DT-$\mathcal{P}_{||}$(*xp_size*) *is in* XP.

Proof. We start by showing the statement of the theorem for DT-LAXP$_{||}$. Let $(\mathcal{T}, e, k)$ be an instance of DT-LAXP$_{||}$. We first enumerate all subsets $A \subseteq F(\mathcal{T})$ of size at most k in time $\mathcal{O}(|F(\mathcal{T})|^k)$. For every such subset A, we then test whether A is a local abductive explanation for e w.r.t. $\mathcal{T}$ in polynomial-time with the help of Lemma 4. If so, we output A as the solution. Otherwise, i.e., if no such subset is a local abductive explanation for e w.r.t. $\mathcal{T}$, we output correctly that $(\mathcal{T}, e, k)$ has no solution.

Let $(\mathcal{T}, c, k)$ be an instance of DT-GAXP$_{||}$. We first enumerate all subsets $A \subseteq F(\mathcal{T})$ of size at most k in time $\mathcal{O}(|F(\mathcal{T})|^k)$. For every such subset A, we then enumerate all of the at most $2^{|A|} \leq 2^k$ partial examples (assignments) $\tau : A \to \{0, 1\}$ in time $\mathcal{O}(2^k)$. For every such partial example τ, we then use Lemma 4 to test whether τ is a global abductive explanation for c w.r.t. $\mathcal{T}$ in polynomial-time. If so, we output e as the solution. Otherwise, i.e., if no such partial example is a global abductive explanation for c w.r.t. $\mathcal{T}$, we output correctly that $(\mathcal{T}, c, k)$ has no solution. The total runtime of the algorithm is at most $2^k|F(\mathcal{T})|^k|\mathcal{T}|^{\mathcal{O}(1)}$.

The algorithm for DT-GCXP$_{||}$ is now very similar to the above algorithm for DT-GAXP$_{||}$. □

The next theorem uses our result that the considered problems are in polynomial-time for DTs (Theorem 5) together with an XP-algorithm that transforms any DT$_{\text{MAJ}}$ into an equivalent DT.

Theorem 7. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{LCXP}, \text{GAXP}, \text{GCXP}\}$. DT$_{\text{MAJ}}$-$\mathcal{P}_{\subseteq}$(*ens_size*) *and* DT$_{\text{MAJ}}$-LCXP$_{||}$(*ens_size*) *are in* XP.

5.3 DSs, DLs and their Ensembles

This subsection is devoted to our algorithmic results for DS, DLs and their ensembles. Our first algorithmic result is again based on our meta-theorem (Theorem 2) and a suitable translation from DS$_{\text{MAJ}}$ and DL$_{\text{MAJ}}$ to a Boolean circuit.

Corollary 8. *Let* $\mathcal{M} \in \{\text{DS}_{\text{MAJ}}, \text{DL}_{\text{MAJ}}\}$ *and let* $\mathcal{P} \in \{\text{LAXP}, \text{LCXP}, \text{GAXP}, \text{GCXP}\}$. $\mathcal{M}$-$\mathcal{P}_{||}$(*ens_size* + *terms_elem*) *is* FPT.

Unlike, DTs, where DT-LCXP$_{||}$ is solvable in polynomial-time, this is not the case for DS-LCXP$_{||}$. Nevertheless, we are able to provide the following result, which shows that DS-LCXP$_{||}$ (and even DL-LCXP$_{||}$) is fixed-parameter tractable parameterized by *term_size* and *xp_size*. The algorithm is based on a novel characterization of local contrastive explanations for DLs.

Lemma 9. *Let* $\mathcal{M} \in \{\text{DS}, \text{DL}\}$. $\mathcal{M}$-LCXP$_{||}$ *for* $M \in \mathcal{M}$ *and integer* k *can be solved in time* $\mathcal{O}(a^k \|M\|^2)$*, where* a *is equal to* term_size*.*

Proof Sketch. Since any DS can be easily translated into a DL without increasing the size of any term, it suffices to show the lemma for DLs. Let (L, e, k) be an instance of DL-LCXP$_{||}$, where $L = (r_1 = (t_1, c_1), \ldots, r_\ell = (t_\ell, c_\ell))$ is a DL, and let r_i be the rule that classifies e, i.e., the first rule that applies to e.

Let R be the set of all rules r_j of L with $c_j \neq c_i$. For a rule $r \in R$, let $A \subseteq F(L)$ such that the example e_A, i.e., the example obtained from e after setting $e_A(f) = 1 - e(f)$ for every $f \in A$ and $e_A(f) = e(f)$ otherwise, is classified by rule r. We claim that:

(1) For every $r \in R$ and every set $A \subseteq F(\mathcal{T})$ such that e_A is classified by r, it holds that A is a local contrastive explanation for e w.r.t. L.

(2) Every local contrastive explanation A for e w.r.t. L contains a subset $A' \subseteq A$ for which there is a rule $r \in R$ such that $e_{A'}$ is classified by r.

Because of (1) and (2), it holds that a set $A \subseteq F(\mathcal{T})$ is a local contrastive explanation if and only if there is a rule $r \in R$ such that e_A is classified by r. Therefore, it is sufficient to be able to compute a minimum set of features A such that e_A is classified by r for every rule $r \in R$, which can be achieved via a bounded-depth branching algorithm. □

The following lemma is now a natural extension of Lemma 9 for ensembles of DLs.

Lemma 10. *Let* $\mathcal{M} \in \{\text{DS}_{\text{MAJ}}, \text{DL}_{\text{MAJ}}\}$. $\mathcal{M}$-LCXP$_{||}$ *for* $M \in \mathcal{M}$ *and integer* k *can be solved in time* $\mathcal{O}(m^s a^k \|M\|^2)$*, where* m *is* terms_elem*,* s *is* ens_size*, and* a *is* term_size*.*

The following theorem now follows immediately from Lemma 10.

Theorem 11. *Let* $\mathcal{M} \in \{\text{DS}, \text{DL}\}$. $\mathcal{M}$-LCXP$_{||}$(*terms_elem* + *xp_size*) *is* FPT*, when ens_size is constant.*

5.4 OBDDs and their Ensembles

In this subsection, we will present our algorithmic results for OBDDs and their ensembles OBDD$^{<}_{\text{MAJ}}$ and OBDD$_{\text{MAJ}}$. Interestingly, while seemingly more powerful OBDDs and OBDD$^{<}_{\text{MAJ}}$s behave very similar to DTs and DT$_{\text{MAJ}}$s if one replaces *mnl_size* with *width_elem*. On the other hand, allowing different orderings for every ensemble OBDD makes OBDD$_{\text{MAJ}}$s much more powerful and harder to explain (see Section 6.3 for an explanation of this phenomenon).

The following two corollaries follow from our meta-theorem Theorem 2 using suitable translations of OBDDs, OBDD$_{\text{MAJ}}$s, and OBDD$^{<}_{\text{MAJ}}$s into Boolean circuits. While it is sufficient to use *width_elem* as a parameter for OBDD$^{<}_{\text{MAJ}}$s; this is no longer the case for OBDD$_{\text{MAJ}}$s, where

one needs to bound the size (instead of the width) of every element in the ensemble.

Corollary 12. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{LCXP}, \text{GAXP}, \text{GCXP}\}$. $\text{OBDD}^{<}_{\text{MAJ}}\text{-}\mathcal{P}_{||}(\textit{ens\_size} + \textit{width\_elem})$ *is* FPT.

Corollary 13. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{LCXP}, \text{GAXP}, \text{GCXP}\}$. $\text{OBDD}_{\text{MAJ}}\text{-}\mathcal{P}_{||}(\textit{ens\_size} + \textit{size\_elem})$ *is* FPT.

The proof of the following theorem is very similar to the corresponding result for DTs (Theorem 5).

Theorem 14. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{LCXP}, \text{GAXP}, \text{GCXP}\}$. OBDD-$\mathcal{P}_{\subseteq}$ *and* OBDD-$\text{LCXP}_{||}$ *can be solved in polynomial-time.*

The next theorem uses our result that the considered problems are in polynomial-time for OBDDs (Theorem 14) together with an XP-algorithm that transforms any $\text{OBDD}^{<}_{\text{MAJ}}$ into an equivalent OBDD.

Theorem 15. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{LCXP}, \text{GAXP}, \text{GCXP}\}$. $\text{OBDD}^{<}_{\text{MAJ}}\text{-}\mathcal{P}_{\subseteq}(\textit{ens\_size})$ *and* $\text{OBDD}^{<}_{\text{MAJ}}\text{-}\text{LCXP}_{||}(\textit{ens\_size})$ *are in* XP.

6 Hardness Results

In this section, we provide our algorithmic lower bounds. We start by showing a close connection between the complexity of all of our explanation problems to the following two problems. As we will see the hardness of finding explanations comes from the hardness of deciding whether or not a given model classifies all examples in the same manner. More specifically, from the HOM problem defined below, which asks whether a given model has an example that is classified differently from the all-zero example, i.e., the example being 0 on every feature. We also need the P-HOM problem, which is a parameterized version of HOM that we use to show parameterized hardness results for deciding the existence of local contrastive explanations.

In the following, let $\mathcal{M}$ be a model type.

$\mathcal{M}$-HOMOGENEOUS (HOM)	
INSTANCE:	A model $M \in \mathcal{M}$.
QUESTION:	Is there an example e such that $M(e) \neq M(e_0)$, where e_0 is the all-zero example?

$\mathcal{M}$-P-HOMOGENEOUS (P-HOM)	
INSTANCE:	A model $M \in \mathcal{M}$ and integer k.
QUESTION:	Is there an example e that sets at most k features to 1 such that $M(e) \neq M(e_0)$, where e_0 is the all-zero example?

The following lemma now shows the connection between HOM and the considered explanation problems.

Lemma 16. *Let* $M \in \mathcal{M}$ *be a model,* e_0 *be the all-zero example, and let* $c = M(e_0)$*. The following problems are equivalent:*

(1) M *is a no-instance of* $\mathcal{M}$-HOM.
(2) The empty set is a solution for the instance (M, e_0) *of* $\mathcal{M}$-$\text{LAXP}_{\subseteq}$.
(3) (M, e_0) *is a no-instance of* $\mathcal{M}$-$\text{LCXP}_{\subseteq}$.
(4) The empty set is a solution for the instance (M, c) *of* $\mathcal{M}$-$\text{GAXP}_{\subseteq}$.
(5) The empty set is a solution for the instance $(M, 1-c)$ *of* $\mathcal{M}$-$\text{GCXP}_{\subseteq}$.
(6) $(M, e_0, 0)$ *is a yes-instance of* $\mathcal{M}$-$\text{LAXP}_{||}$.
(7) (M, e_0) *is a no-instance of* $\mathcal{M}$-$\text{LCXP}_{||}$.
(8) $(M, c, 0)$ *is a yes-instance of* $\mathcal{M}$-$\text{GAXP}_{||}$.
(9) $(M, 1-c, 0)$ *is a yes-instance of* $\mathcal{M}$-$\text{GCXP}_{||}$.

Proof. It is easy to verify that all of the statements (1)–(9) are equivalent to the following statement (and therefore equivalent to each other): $M(e) = M(e_0) = c$ for every example e. □

While Lemma 16 is sufficient for most of our hardness results, we also need the following lemma to show certain parameterized hardness results for deciding the existence of local contrastive explanations.

Lemma 17. *Let* $M \in \mathcal{M}$ *be a model and let* e_0 *be the all-zero example. The following problems are equivalent:*

(1) (M, k) *is a yes-instance of* $\mathcal{M}$-P-HOM.
(2) (M, e_0, k) *is a yes-instance of* $\mathcal{M}$-$\text{LCXP}_{||}$.

We will often reduce from the following problem, which is well-known to be NP-hard and also W[1]-hard parameterized by k.

MULTICOLORED CLIQUE (MCC)	
INSTANCE:	A graph G with a proper k-coloring of $V(G)$.
QUESTION:	Is there a clique of size k in G?

The following lemma provides a unified way to show hardness results for ensembles for practically all of our model types in the case that we allow arbitrarily many (constant-size) ensemble elements, i.e., we use it to show Theorems 24, 28, 31.

Lemma 18. *Let* $\mathcal{M}$ *be a class of models such that there are models* $M^0 \in \mathcal{M}$, $M^1_f \in \mathcal{M}$ *and* $M^2_{f_1,f_2} \in \mathcal{M}$ *for features* f, f_1, *and* f_2 *of size at most* d *such that:*

- M^0 *classifies every example negatively.*
- M^1_f *classifies an example* e *positively iff* $e(f) = 1$.
- $M^2_{f_1,f_2}$ *classifies an example* e *positively iff* $e(f_1) = 0$ *or* $e(f_2) = 0$.

$\mathcal{M}_{\text{MAJ}}$-P-HOM *is* W[1]*-hard parameterized by* k *even if the size of each ensemble element is at most* d *and* $\mathcal{M}_{\text{MAJ}}$-HOM *is* NP*-hard even if the size of each ensemble element is at most* d*.*

Proof Sketch. We provide a parameterized reduction from the MULTICOLORED CLIQUE (MCC) problem, which is also a polynomial-time reduction. Given an instance (G, k) of the MCC problem with k-partition $(V_1, \ldots, V_k)$ of $V(G)$, we will construct an equivalent instance $(\mathcal{E}, k)$ of $\mathcal{M}_{\text{MAJ}}$-P-HOM in polynomial-time as follows. $\mathcal{E}$ uses one binary feature f_v for every $v \in V(G)$. Let $<_V$ be an arbitrary ordering of $V(G)$. We denote by n and m the number of vertices and edges of the graph G, respectively.

$\mathcal{E}$ contains the following ensemble elements:

- For every non-edge $uv \notin E(G)$ with $u <_V v$, we add the model $M^2_{f_u,f_v}$ to $\mathcal{E}$.
- For every vertex $v \in V(G)$, we add the model $M^1_{f_v}$ to $\mathcal{E}$.
- We add $(\binom{n}{2} - m) - n + 2k - 1$ models M^0 to $\mathcal{E}$.

Clearly, the reduction works in polynomial-time and preserves the parameter and it only remains to show that G has a k-clique if and only if there is an example e such that $\mathcal{E}(e) \neq \mathcal{E}(e_0)$ that sets at most k features to 1. □

6.1 DTs and their Ensembles

Here, we provide our algorithmic lower bounds for DTs and their ensembles. We say that a DT $\mathcal{T}$ is *ordered* if there is an ordering $<$ of the features in $F(\mathcal{T})$ such that the ordering of the features on every root-to-leaf path of $\mathcal{T}$ agrees with $<$. We need the following auxiliary lemma to simplify the descriptions of our reductions.

Lemma 19. *Let $E \subseteq E(F)$ be a set of examples defined on features in F. An ordered* DT *$\mathcal{T}_E$ of size at most $2|E||F|+1$ such that $\mathcal{T}_E(e) = 1$ if and only if $e \in E$ can be constructed in time $\mathcal{O}(|E||F|)$.*

Proof. Let $< = (f_1, \ldots, f_n)$ be an arbitrary order of the features in F. First, we construct a simple ordered DT $\mathcal{T}_e = (T_e, \lambda_e)$ that classifies only example e as 1 and all other examples as 0. $\mathcal{T}_e$ has one inner node t^e_i for every $i \in [n]$ with $\lambda_e(t^e_i) = f_i$. Moreover, for $i < n$, t^e_i has t^e_{i+1} as its $e(f_i)$-child and a new 0-leaf as its other child. Finally, t^e_n has a new 1-leaf as its $e(f_n)$-child and a 0-leaf as its other child. Clearly, $\mathcal{T}_e$ can be constructed in time $\mathcal{O}(|F|)$.

We now construct $\mathcal{T}_E$ iteratively starting from $\mathcal{T}_\emptyset$ and adding one example from E at a time (in an arbitrary order). We set $\mathcal{T}_\emptyset$ to be the DT that only consists of a 0-leaf. Now to obtain $\mathcal{T}_{E' \cup \{e\}}$ from $\mathcal{T}_{E'}$ for some $E' \subseteq E$ and $e \in E \setminus E'$, we do the following. Let l be the 0-leaf of $\mathcal{T}_{E'}$ that classifies e and let f_i be the feature assigned to the parent of l. Moreover, let $\mathcal{T}'_e$ be the sub-DT of $\mathcal{T}_e$ rooted at t^e_{i+1} or if $i = n$ let $\mathcal{T}'_e$ be the DT consisting only of a 1-leaf. Then, $\mathcal{T}_{E' \cup \{e\}}$ is obtained from the disjoint union of $\mathcal{T}_{E'}$ and $\mathcal{T}'_e$ after identifying the root of $\mathcal{T}'_e$ with l. Clearly, $\mathcal{T}_E$ is an ordered DT that can be constructed in time $\mathcal{O}(|E||F|)$ has size at most $2|E||F| + 1$ and satisfies $\mathcal{T}_E(e) = 1$ if and only if $e \in E$. □

We note that the following theorem also follows from a result in (Barceló et al. 2020, Proposition 5) for FBDDs, i.e., BDDs without contradicting paths. However, we require a different version of the proof that generalizes easily to OBDDs, i.e., we need to show hardness for ordered DTs.

Theorem 20. DT-LAXP$_{||}$ *is* NP*-hard and* DT-LAXP$_{||}$(*xp_size*) *is* W[2]*-hard even if for ordered* DT*s.*

The following theorem is an analogue of Theorem 20 for global abductive and global contrastive explanations. It is interesting to note that while it was not necessary to distinguish between local abductive explanations on one side and global abductive and global contrastive explanations on the other side in the setting of algorithms, this is no longer the case when it comes to algorithmic lower bounds. Moreover, while the following result establishes W[1]-hardness for DT-GAXP$_{||}$(*xp_size*) and DT-GCXP$_{||}$(*xp_size*), this is achieved via fpt-reductions that are not polynomial-time reductions, which is a behavior that is very rarely seen in natural parameterized problems. While it is therefore not clear whether the problems are NP-hard, the result still shows that the problems are not solvable in polynomial-time unless FPT = W[1], which is considered unlikely (Downey and Fellows 2013).

Theorem 21. DT-GAXP$_{||}$(*xp_size*) *and* DT-GCXP$_{||}$(*xp_size*) *are* W[1]*-hard. Moreover, there is no polynomial time algorithm for solving* DT-GAXP$_{||}$ *and* DT-GCXP$_{||}$*, unless* FPT = W[1].

Proof Sketch. We provide a parameterized reduction from the MULTICOLORED CLIQUE (MCC) problem, which is well-known to be W[1]-hard parameterized by the size of the solution. Given an instance (G, k) of the MCC problem with k-partition $(V_1, \ldots, V_k)$ of $V(G)$, we will construct an equivalent instance $(\mathcal{T}, c, k)$ of GAXP$_{||}$ in fpt-time. Note that since a partial example e' is a global abductive explanation for c w.r.t. $\mathcal{T}$ if and only if e' is a global contrastive explanation for $1 - c$ w.r.t. $\mathcal{T}$, this then also implies the statement of the theorem for GCXP$_{||}$. $\mathcal{T}$ uses one binary feature f_v for every $v \in V(G)$.

We start by constructing the DT $\mathcal{T}_{i,j}$ for every $i, j \in [k]$ with $i \neq j$ satisfying the following: (*) $\mathcal{T}_{i,j}(e) = 1$ for an example e if and only if either $e(f_v) = 0$ for every $v \in V_i$ or there exists $v \in V_i$ such that $e(f_v) = 1$ and $e(f_{v'}) = 0$ for every $v' \in (V_i \setminus \{v\}) \cup (N_G(v) \cap V_j)$. Let $\mathcal{T}_i$ be the DT obtained using Lemma 19 for the set of examples $\{e_0\} \cup \{ e_v \mid v \in V_i \}$ defined on the features in $F_i = \{ f_v \mid v \in V_i \}$. Here, e_0 is the all-zero example and for every $v \in V_i$, e_v is the example that is 1 only at the feature f_v and 0 otherwise. Moreover, for every $v \in V_i$, let $\mathcal{T}^v_j$ be the DT obtained using Lemma 19 for the set of examples containing only the all-zero example defined on the features in $\{ f_{v'} \mid v' \in N_G(v) \cap V_j \}$. Then, $\mathcal{T}_{i,j}$ is obtained from $\mathcal{T}_i$ after replacing the 1-leaf that classifies e_v with $\mathcal{T}^v_j$ for every $v \in V_i$. Clearly, $\mathcal{T}_{i,j}$ satisfies (*) and since $\mathcal{T}_i$ has at most $|V_i|^2$ inner nodes and $\mathcal{T}^v_j$ has at most $|V_j|$ inner nodes, we obtain that $\mathcal{T}_{i,j}$ has at most $\mathcal{O}(|V(G)|^2)$ nodes.

For an integer ℓ, we denote by DT(ℓ) the complete DT of height ℓ, where every inner node is assigned to a fresh auxiliary feature and every of the exactly 2^ℓ leaves is a 0-leaf. Let $\mathcal{T}_\Delta$ be the DT obtained from the disjoint union of $\mathcal{T}_U = \text{DL}(k)$ and 2^k copies $\mathcal{T}^1_D, \ldots, \mathcal{T}^{2^k}_D$ of DT ($\lceil \log(k(k-1)) \rceil$) by identifying the i-th leaf of $\mathcal{T}_U$ with the root of $\mathcal{T}^i_D$ for every i with $1 \leq i \leq 2^k$; each copy is equipped with its own set of fresh features.

Then, $\mathcal{T}$ is obtained from $\mathcal{T}_\Delta$ after doing the following with $\mathcal{T}^\ell_D$ for every $\ell \in [2^k]$. For every $i, j \in [k]$ with $i \neq j$, we replace a private leaf of $\mathcal{T}^\ell_D$ with the DT $\mathcal{T}_{i,j}$; note that this is possible because $\mathcal{T}^\ell_D$ has at least $k(k-1)$ leaves. Also note that $\mathcal{T}$ has size at most $\mathcal{O}(|\mathcal{T}_\Delta||V(G)|^2)$. This completes the construction of $\mathcal{T}$ and we set $c = 0$. Clearly, $\mathcal{T}$ can be constructed from G in fpt-time w.r.t. k. It remains

to show that G has a k-clique if and only if there is a global abductive explanation of size at most k for c w.r.t. $\mathcal{T}$. □

Lemma 22. DT_{MAJ}-HOM *is* NP*-hard and both* DT_{MAJ}-HOM(*ens_size*) *and* DT_{MAJ}-P-HOM(*ens_size*) *are* W[1]*-hard.*

Proof Sketch. We give a parameterized reduction from MCC that is also a polynomial-time reduction. That is, given an instance (G, k) of MCC with k-partition $V_1, \ldots, V_k$, we will construct a DT_{MAJ} $\mathcal{E}$ with $|\mathcal{E}| = 2(k + \binom{k}{2}) - 1$ such that G has a k-clique if and only if $\mathcal{E}$ classifies at least one example positively. This will already suffice to show the stated results for DT_{MAJ}-HOM. Moreover, to show the results for DT_{MAJ}-P-HOM we additional show that G has a k-clique if and only if $\mathcal{E}$ classifies an example positively that sets at most k features to 1.

$\mathcal{E}$ will use the set of features $\bigcup_{i \in [k]} F_i$, where $F_i = \{ f_v \mid v \in V_i \}$. For each $v \in V_i$ and $u \in V_j$, let $e_{v,u}$ be an example defined on set of features $F_i \cup F_j$ that is 1 only at the features f_v and f_v, and otherwise 0. For every $i \in [k]$, $\mathcal{E}$ will have a DT $\mathcal{T}_i$ obtained using Lemma 19 for the set of examples $\{ e_{v,v} \mid v \in V_i \}$ defined on the features in F_i. Also, for every i and j with $1 \leq i < j \leq k$, $\mathcal{E}$ contains a DT $\mathcal{T}_{i,j}$ obtained using Lemma 19 for the set of examples $\{ e_{v,u} \mid v \in V_i \wedge u \in V_j \wedge vu \in E(G) \}$ defined on the features in $F_i \cup F_j$. Finally, $\mathcal{E}$ contains $k + \binom{k}{2} - 1$ DTs that classify every example negatively, i.e., those DTs consists only of one 0-leaf. The correctness of the reduction is provided in the long version of the paper. □

The final two theorems of this section provide all the remaining hardness results for DT_{MAJ}s and follow from Lemma 22 and Lemma 18, respectively, together with Lemmas 16, 17.

Theorem 23. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. DT_{MAJ}-$\mathcal{P}_{\subseteq}$(*ens_size*) *is* co-W[1]*-hard;* DT_{MAJ}-LCXP$_{\subseteq}$(*ens_size*) *is* W[1]*-hard;* DT_{MAJ}-$\mathcal{P}_{||}$(*ens_size*) *is* co-W[1]*-hard even if xp_size is constant;* DT_{MAJ}-LCXP$_{||}$(*ens_size* + *xp_size*) *is* W[1]*-hard.*

Theorem 24. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. DT_{MAJ}-$\mathcal{P}_{\subseteq}$ *is* co-NP*-hard even if mnl_size + size_elem is constant;* DT_{MAJ}-LCXP$_{\subseteq}$ *is* NP*-hard even if mnl_size + size_elem is constant;* DT_{MAJ}-$\mathcal{P}_{||}$ *is* co-NP*-hard even if mnl_size + size_elem + xp_size is constant;* DT_{MAJ}-LCXP$_{||}$(*xp_size*) *is* W[1]*-hard even if mnl_size + size_elem is constant.*

6.2 DSs, DLs and their Ensembles

Here, we establish our hardness results for DS, DLs, and their ensembles. It is interesting to note that there is no real distinction between DS and DLs when it comes to explainability and that both are considerably harder to explain than DTs and OBDDs.

Theorem 25. *Let* $\mathcal{M} \in \{\text{DS}, \text{DL}\}$ *and let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. $\mathcal{M}$-$\mathcal{P}_{\subseteq}$ *is* co-NP*-hard even if term_size is constant;* $\mathcal{M}$-LCXP$_{\subseteq}$ *is* NP*-hard even if term_size is constant;* $\mathcal{M}$-$\mathcal{P}_{||}$ *is* co-NP*-hard even if term_size + xp_size is constant.*

Theorem 26. *Let* $\mathcal{M} \in \{\text{DS}, \text{DL}\}$. $\mathcal{M}$-LCXP$_{||}$(*xp_size*) *is* W[1]*-hard.*

Theorem 27. *Let* $\mathcal{M} \in \{\text{DS}_{\text{MAJ}}, \text{DL}_{\text{MAJ}}\}$. $\mathcal{M}$-LCXP$_{||}$(*ens_size* + *xp_size*) *is* W[1]*-hard even if term_size is constant.*

Theorem 28. *Let* $\mathcal{M} \in \{\text{DS}_{\text{MAJ}}, \text{DL}_{\text{MAJ}}\}$ *and let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. $\mathcal{M}$-$\mathcal{P}_{\subseteq}$ *is* co-NP*-hard even if terms_elem + term_size is constant;* $\mathcal{M}$-LCXP$_{\subseteq}$ *is* NP*-hard even if terms_elem + term_size is constant;* $\mathcal{M}$-$\mathcal{P}_{||}$ *is* co-NP*-hard even if terms_elem + term_size + xp_size is constant;* $\mathcal{M}$-LCXP$_{||}$(*xp_size*) *is* W[1]*-hard even if terms_elem + term_size is constant.*

6.3 OBDDs and their Ensembles

We are now ready to provide our hardness results for OBDDs and their ensembles $\text{OBDD}^{<}_{\text{MAJ}}$s and OBDD_{MAJ}s. While the proofs for OBDDs and $\text{OBDD}^{<}_{\text{MAJ}}$s follow along very similar lines as the corresponding proofs for DTs, the main novelty and challenge of this subsection are the much stronger hardness results for OBDD_{MAJ}s. Informally, we show that the satisfiability of any CNF formula ϕ can be modelled in terms of an ensemble of two OBDDs O_1 and O_2 each using a different ordering of the variables. In particular, it holds that both OBDDs classify an example positively if and only if the corresponding assignment satisfies ϕ. The main idea behind the construction of O_1 and O_2 is to make copies for every occurrence of a variable in ϕ and then use O_1 to verify that the assignment satisfies ϕ and O_2 to verify that all copies of every variable are assigned to the same value.

Theorem 29. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. OBDD_{MAJ}-$\mathcal{P}_{\subseteq}$ *is* co-NP*-hard even if ens_size + width_elem is constant;* OBDD_{MAJ}-LCXP$_{\subseteq}$ *is* NP*-hard even if ens_size + width_elem is constant;* OBDD_{MAJ}-$\mathcal{P}_{||}$ *is* co-NP*-hard even if ens_size + width_elem + xp_size is constant;* OBDD_{MAJ}-LCXP$_{||}$(*ens_size*) *is* W[1]*-hard even if ens_size + width_elem is constant.*

A special case of the following theorem, the NP-hardness of LAXP$_{||}$ for FBDDs, i.e., "free" BDDs, was shown by Barceló et al. (2020).

Theorem 30. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. OBDD-$\mathcal{P}_{||}$ *is* NP*-hard and* OBDD-$\mathcal{P}_{||}$(*xp_size*) *is* W[2]*-hard.*

Theorem 31. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. $\text{OBDD}^{<}_{\text{MAJ}}$-$\mathcal{P}_{\subseteq}$ *is* co-NP*-hard even if width_elem + size_elem is constant;* $\text{OBDD}^{<}_{\text{MAJ}}$-LCXP$_{\subseteq}$ *is* NP*-hard even if width_elem + size_elem is constant;* $\text{OBDD}^{<}_{\text{MAJ}}$-$\mathcal{P}_{||}$ *is* co-NP*-hard even if width_elem + size_elem + xp_size is constant;* $\text{OBDD}^{<}_{\text{MAJ}}$-LCXP$_{||}$(*xp_size*) *is* W[1]*-hard even if width_elem + size_elem is constant;*

Theorem 32. *Let* $\mathcal{P} \in \{\text{LAXP}, \text{GAXP}, \text{GCXP}\}$. $\text{OBDD}^{<}_{\text{MAJ}}$-$\mathcal{P}_{\subseteq}$(*ens_size*) *is* co-W[1]*-hard;* $\text{OBDD}^{<}_{\text{MAJ}}$-LCXP$_{\subseteq}$(*ens_size*) *is* W[1]*-hard;* $\text{OBDD}^{<}_{\text{MAJ}}$-$\mathcal{P}_{||}$(*ens_size*) *is* co-W[1]*-hard even if xp_size is constant;* $\text{OBDD}^{<}_{\text{MAJ}}$-LCXP$_{||}$(*ens_size* + *xp_size*) *is* W[1]*-hard.*

7 Conclusion

We have developed an in-depth exploration of the parameterized complexity of explanation problems in various machine learning (ML) models, focusing on models with transparent internal mechanisms. By analyzing different models and their ensembles, we have provided a comprehensive overview of the complexity of finding explanations in these systems. These insights are crucial for understanding the inherent complexity of different ML models and their implications for explainability.

Among our findings, some results stand out as particularly unexpected. For instance, while $\mathrm{DT}_{\mathrm{MAJ}}$ and $\mathrm{OBDD}^{<}_{\mathrm{MAJ}}$s are seemingly different model types, our results show that they behave similarly w.r.t. tractability for explanation problems. On the other hand, it seems surprising that many of the tractability results that hold for DTs and OBDDs do not carry over to seemingly simpler models such as DSs and DLs. For instance, while all variants of LCXP are polynomial-time for DTs and OBDDs, this is not the case for DSs or DLs. Nevertheless, we obtain interesting FPT-algorithms for DL-$\mathrm{LCXP}_{||}$ (Theorem 11). $\mathrm{OBDD}_{\mathrm{MAJ}}$ stands out as the hardest model for computing explanations by far, which holds even for models with only two ensemble elements. From a complexity point of view, DT-$\mathrm{GAXP}_{||}$ provides the rare scenario where a problem is known as W[1]-hard but not confirmed to be NP-hard (Theorem 21).

Looking ahead, there are several promising directions for future research. First, we aim to extend our complexity classification to Sequential Decision Diagrams (Darwiche 2011) or even FBDDs, which offer a more succinct representation than OBDDs (Bova 2016). This extension could provide further insights into the complexity of explanations in more compact ML models. Secondly, we propose to explore other problem variations, such as counting different types of explanations or finding explanations that meet specific constraints beyond just the minimum ones (Barceló et al. 2020). Lastly, the concept of weighted ensembles presents an intriguing avenue for research. While the hardness results we established likely still apply, the tractability in the context of weighted ensembles needs to be clarified and warrants further investigation. It would be interesting to see how our results hold up when considering polynomial-sized weights.

Acknowledgements

Stefan Szeider acknowledges support by the Austrian Science Fund (FWF) within the projects 10.55776/P36688, 10.55776/P36420, and 10.55776/COE12. Sebastian Ordyniak was supported by the Engineering and Physical Sciences Research Council (EPSRC) (Project EP/V00252X/1).

References

Barceló, P.; Monet, M.; Pérez, J.; and Subercaseaux, B. 2020. Model interpretability through the lens of computational complexity. *Proc. NeurIPS 2020* 33:15487–15498.

Bergougnoux, B.; Dreier, J.; and Jaffke, L. 2023. A logic-based algorithmic meta-theorem for mim-width. *Proc. SODA 2023* 3282–3304.

Bova, S. 2016. SDDs are exponentially more succinct than OBDDs. *Proc. AAAI 2016* 929–935.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8(8).

Chan, H., and Darwiche, A. 2003. Reasoning about bayesian network classifiers. *Proc. UAI 2003* 107–115.

Commission, E. 2019. *Ethics guidelines for trustworthy AI*. Publications Office, European Commission and Directorate-General for Communications Networks, Content and Technology.

Commission, E. 2020. *White Paper on Artificial Intelligence: a European approach to excellence and trust*. Publications Office, European Commission and Directorate-General for Communications Networks, Content and Technology.

Darwiche, A., and Ji, C. 2022. On the computation of necessary and sufficient explanations. *Proc. AAAI 2022* 5582–5591.

Darwiche, A. 2011. SDD: A new canonical representation of propositional knowledge bases. *Proc. IJCAI 2011* 819–826.

Downey, R. G., and Fellows, M. R. 2013. *Fundamentals of parameterized complexity*. Texts in Computer Science. Springer Verlag.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys* 51(5):93:1–93:42.

Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; and Samek, W. 2020. Explainable AI methods - A brief overview. *Proc. xxAI@ICML 2020* 13200:13–38.

Ignatiev, A.; Narodytska, N.; NicholasAsher; and Marques-Silva, J. 2020. From contrastive to abductive explanations and back again. *Proc. AIxIA 2020* 12414:335–355.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-based explanations for machine learning models. *Proc. AAAI 2019* 1511–1519.

Lipton, Z. C. 2018. The mythos of model interpretability. *Communications of the ACM* 61(10):36–43.

Lisboa, P. J. G.; Saralajew, S.; Vellido, A.; Fernández-Domenech, R.; and Villmann, T. 2023. The coming of age of interpretable and explainable machine learning models. *Neurocomputing* 535:25–39.

Marques-Silva, J. 2023. Logic-based explainability in machine learning. *Reasoning Web. Causality, Explanations and Declarative Knowledge* 24–104.

Mengel, S., and Slivovsky, F. 2021. Proof complexity of symbolic QBF reasoning. *Proc. SAT* 12831:399–416.

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

Molnar, C. 2023. *Interpretable Machine Learning*. Lulu.com.

OECD. 2023. The state of implementation of the OECD AI principles four years on. Technical Report 3, The Organisation for Economic Co-operation and Development.

Ordyniak, S.; Paesani, G.; Rychlicki, M.; and Szeider, S. 2024. Explaining decisions in ML models: a parameterized complexity analysis. *arXiv* (2407.15780).

Ordyniak, S.; Paesani, G.; and Szeider, S. 2023. The parameterized complexity of finding concise local explanations. *Proc. IJCAI 2023* 3312–3320.

Oum, S., and Seymour, P. D. 2006. Approximating clique-width and branch-width. *Journal of Combinatorial Theory* 96(4):514–528.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should I trust you?": Explaining the predictions of any classifier. *Proc. KDD 2016* 1135–1144.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining Bayesian network classifiers. *Proc. IJCAI 2018* 5103–5111.