Computing Probabilistic Explanations for ML Models: Fixed-Parameter Algorithms

Sebastian Ordyniak[1], **Mateusz Rychlicki**[1], **Stefan Szeider**[2]

[1]School of Computer Science, University of Leeds, UK
[2]Algorithms and Complexity Group, TU Wien, Austria
sordyniak@gmail.com, mkrychlicki@gmail.com, stefan@szeider.net

Abstract

Machine learning models now drive many critical decisions, making explanations of their reasoning essential. Recent work analyzes the complexity of exact explanations in transparent models, but these explanations are often too large for practical use. This has motivated research into probabilistic alternatives.

We study probabilistic extensions that allow controlled uncertainty while maintaining rigorous foundations. We analyze three basic model types: decision trees, decision lists, and decision sets. We introduce algorithms for computing both local and global probabilistic explanations for these models. Our main result shows that computing minimum-size probabilistic explanations is fixed-parameter tractable when parameterized by structural properties—specifically, the number of terms for decision lists and decision sets and the minimum of the number of positive and the number of negative leaves.

1 Introduction

Machine learning models make automated decisions that impact human lives, creating a need for formal methods to explain these decisions (Ignatiev, Narodytska, and Marques-Silva 2019; Guidotti et al. 2019). For models with accessible internal mechanisms, such as decision trees (DTs), decision sets (DSs), and decision lists (DLs), these explanations can be computed directly from the model structure instead of treating them as black boxes (Barceló et al. 2020).

We consider these fundamental types of formal explanations:

Local abductive explanations (Ignatiev, Narodytska, and Marques-Silva 2019; Marques-Silva 2023), also known as sufficient reason explanations (Darwiche and Ji 2022), identify a minimal subset of the set F of all features that maintain the model's prediction. Given an input example x classified as c, such an explanation is a subset of F where fixing their values from x forces the classification c regardless of other feature values.

Global abductive explanations (Ribeiro, Singh, and Guestrin 2016), also referred to as prime-implicant explanations (Shih, Choi, and Darwiche 2018), provide again a subset of F such that a valuation of the features in the subset guarantees a specific classification c. They identify feature value combinations that force the model to output c for any completion of the remaining features.

Global contrastive explanations (Miller 2019; Ignatiev et al. 2020), known as necessary reason explanations (Darwiche and Ji 2022), are dual to global abductive explanations and identify a minimal subset of features that prevent a specific classification c.

These explanations often grow large and can exceed human cognitive limits for understanding model decisions (Marques-Silva 2023).

To address these challenge, one can consider *probabilistic versions* of the above formal explanation types (Izza et al. 2023), which maintain rigorous foundations while allowing for controlled uncertainty.

For instance, in a *probabilistic local abductive explanation*, instead of requiring all completions of the remaining features to maintain the original classification, we only require that a fraction γ of these completions do so. This threshold γ allows for smaller explanations by tolerating some exceptions, creating a trade-off between explanation size and classification certainty. For the two global explanation types, one can define probabilistic variants similarly[1].

Setting $\gamma = 1$ in probabilistic explanations recovers the non-probabilistic versions, hence the non-probabilistic variants are indeed a special case of the probabilistic ones.

Finding smallest non-probabilistic explanations is computationally hard: All variants we consider are NP-hard or coNP-hard for DTs, DSs, and DLs (Ordyniak et al. 2024).

Motivated by the intractability in the classical complexity setting, Ordyniak et al. (2024) identified natural parameterizations that yield fixed-parameter tractability[2]. For DTs, the smallest number of leaves with the same classification turns out to be the critical parameter, while for DSs and DLs, it is the number of terms that matters. While Ordyniak et al. (2024) provides a comprehensive picture of the parameterized complexity landscape of computing non-probabilitic explanations for a wide variety of models such as DTs, DSs,

[1]Local contrastive explanations identify a minimal feature change that flips the model's prediction for a given input x and has no meaningful probabilistic variant.

[2]The notion of fixed-parameter tractability strictly extends polytime tractability by allowing a constant factor that is possibly exponential in a problem parameter while keeping the running time polynomial in the input size (Downey and Fellows 2013).

DLs, OBDD, BDD and even ensembles thereof, the corresponding question for the much more general (and arguably much more practically relevant) probabilistic explanations is thus far entirely unexplored.

In this work we initiate the parameterized complexity analysis of probabilistic explanations. As our main result, we provide a unified algorithmic framework that allows us to generalise all of the recently shown fixed-parameter tractability results for the fundamental models of DTs, DSs, and DLs from the non-probabilistic to the probabilistic setting (Ordyniak et al. 2024). This is rather surprising given that the probabilistic setting can be seen as the counting version of the non-probabilistic setting and is therefore conceptually much harder. Indeed, our results require entirely different and novel algorithmic techniques since the existing techniques based on powerful algorithmic meta-theorems as well as tailor-made dynamic programming algorithms fall well short of providing solutions for the probabilistic setting. As a side result, we are able to improve the runtime of the known algorithms for the non-probabilistic setting significantly.

Our main technical contribution is a unified algorithm that establishes fixed-parameter tractability for computing minimum-size explanations across DTs, DSs, and DLs in both the probabilistic setting and for explanations with *outliers*. In the latter case, instead of requiring exact matches for classifications, we allow for a bounded number δ of misclassifications among the completions.

The key to our unified approach is the development of *"featuristic functions"* $H_T(e)$ that capture the relevant properties of the different model types in a unified manner. Informally, the featuristic function provides a compact representation of the model tailor-made for answering explanation queries and we show that given a model M of any of the considered types $\mathcal{M} \in \{\text{DT}, \text{DL}, \text{DS}\}$ and a classification c, we can efficiently compute a featuristic function $H_T(e)$ that for every partial example e returns the number of examples e' such that e' agrees with e and $M(e) = c$. This allows us to provide a unified approach for all considered model and explanation types.

Statements whose full proofs are omitted due to space constraints are marked with $\star$ and can be found in the full version of the paper.

2 Preliminaries

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 $F(e)$ 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 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 negative example (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 results easily carry over to the classification into any finite set of classes.

Decision Trees. A *decision tree* (DT) $\mathcal{T}$ is binary classification model defined by a pair (T, λ) where 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 for an example $e \in E(F)$ as follows. 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 $b \in \{0,1\}$. 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 $z \in \{0,1\}$. We denote by $F(t)$ the set of features corresponding to literals in t. 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 number of terms of S. 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 $\ell \geq 0$. We denote by $|L|$ the number of rules of L. 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.

3 Probabilistic Explanations

Let F be a set of binary features, $e \in E(F)$, and $A \subseteq F$. We denote by $e[A]$ the partial example obtained from e by restricting the set of features to A. We denote by $E(F, e')$, or briefly $E(e')$ if F is clear from the context, the set of all examples in $E(F)$ that agree with the partial example e'. Moreover, for a model $M : E(F) \to \{0,1\}$ and a class $c \in \{0,1\}$, we denote by $E_M(F, e', c)$, or briefly $E_M(e', c)$ if F is clear from the context, the subset of $E(F, e')$ restricted to examples e'' with $M(e'') = c$.

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

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 $\{z\}$ is the only probabilistic local abductive explanation for $\gamma = 3/4$ and e in L of size at most 1. Let $\tau_1 = \{x \mapsto 1\}$ and $\tau_2 = \{z \mapsto 0\}$ be partial examples. Note that τ_1 and τ_2 are minimal global abductive and global contrastive explanation with outliers for $\delta = 1$ and class 0 w.r.t. L, respectively.

A *probabilistic local abductive explanation (*LAXP_P*) for e w.r.t. M and threshold* $\gamma \in [0, 1]$ is a subset $A \subseteq F(M)$ of features such that: $|E_M(e[A], M(e))|/|E(e[A])| \geq \gamma$.

A *probabilistic global abductive explanation (*GAXP_P*) for c w.r.t. M and threshold* $\gamma \in [0, 1]$ is a partial example e' such that: $|E_M(e', c)|/|E(e')| \geq \gamma$.

A *probabilistic global contrastive explanation (*GCXP_P*) for c w.r.t. M and threshold* $\gamma \in [0, 1]$ is a partial example e' such that: $|E_M(e', 1 - c)|/|E(e')| \geq \gamma$.

We also define a version of each of the above explanation types, which we refer to by "with outliers", that allows to bound the number of outliers or mistakes allowed by the explanation. For instance, the version of LAXP_P with outliers (LAXP_O) is given below and the corresponding versions GAXP_O and GCXP_O of GAXP_P and GCXP_P are defined analogously.

A *local abductive explanation with outliers (*LAXP_O*) for e w.r.t. M and $\delta \in \mathbb{Z}$* is a subset $A \subseteq F(M)$ of features such that: $|E(e[A])| - |E_M(e[A], M(e))| \leq \delta$.

Problems For each of the above explanation types and each of the considered model types $\mathcal{M}$, one can now define the corresponding computational problem. For instance:

$\mathcal{M}$-PROBABILISTIC LOCAL ABDUCTIVE EXPLANATION (LAXP_P)	
INSTANCE:	A model $M \in \mathcal{M}$, an example e, and a threshold γ.
QUESTION:	Find a smallest probabilistic local abductive explanation for e w.r.t. M and γ.

The problems $\mathcal{M}$-X for $X \in \{\mathrm{GAXP}_P, \mathrm{GCXP}_P\}$ are defined analogously.

$\mathcal{M}$-LOCAL ABDUCTIVE EXPLANATION WITH OUTLIERS (LAXP_O)	
INSTANCE:	A model $M \in \mathcal{M}$, an example e, and a number δ.
QUESTION:	Find a smallest local abductive explanation for e w.r.t. M and δ.

Again, the problems $\mathcal{M}$-X for $X \in \{\mathrm{GAXP}_O, \mathrm{GCXP}_O\}$ are defined analogously.

4 Parameters and Overview of Results

The probabilistic explanation problems introduced in the previous section are all NP-hard or co-NP-hard for all considered model types. This follows form the fact that they are at least as hard as their deterministic counterparts that are known to be hard (Ordyniak et al. 2024). Given these hardness results, we need a more refined complexity analysis to identify tractable cases. This naturally leads us to study the parameterized complexity of these problems.

Parameterized complexity analysis has become a standard in many areas, especially in AI, where many of the considered problems are (co)-NP-hard, and one tries to find efficient algorithms whose running time is polynomial in the input size n but depends exponentially on one or several parameters.

The central notion is *fixed-parameter tractability (FPT)* which refers to problems that can be solved in time $f(k)n^{\mathcal{O}(1)}$ for instances of input size n and parameter k. A problem is XP-tractable if it can be solved in time $\mathcal{O}(n^{f(k)})$. The notion of $W[i]$ hardness allows to separate FPT from XP problems. A problem is para-NP-hard if it is NP-hard for the parameter considered constant. We have the following hierarchy of complexity classes, where all inclusions are believed to be strict.

$$P \subseteq \text{FPT} \subseteq W[1] \subseteq W[2] \subseteq \cdots \text{XP} \cap \text{NP} \subseteq \text{para-NP}$$

We refer to the textbook by Downey and Fellows (2013) for fundamentals on parameterized complexity and to recent papers studying the parameterized complexity of problems that arise in explainable AI (Dabrowski et al. 2024; Ordyniak et al. 2024,?; Eiben et al. 2023; Ordyniak, Paesani, and Szeider 2023; Ordyniak and Szeider 2021; Barceló et al. 2020,?).

We provide a comprehensive analysis of the parameterized complexity of all considered problems with respect to all combinations of parameters considered in (Ordyniak et al. 2024), i.e., minimum of the number of positive and the number of negative leaves for DTs (*mnl_size*), largest size of a term for DSs and DLs (*term_size*), number of terms for DSs and DLs (*terms_elem*), and size of the explanation *xp_size*.

As it turns out (Ordyniak et al. 2024) excluded tractability results for all combinations of parameters apart from *mnl_size*, *terms_elem*, and *xp_size* and we show that surprisingly all tractability results given in the deterministic setting carry over to the probabilistic setting. In particular, we show that:

(1) DT-X_Y is FPT parameterized by *mnl_size* for all $X \in \{$ LAXP, GAXP, GCXP$\}$ and $Y \in \{P, O\}$.

(2) DS-X_Y and DL-X_Y are FPT parameterized by *terms_elem* for all $X \in \{$ LAXP, GAXP, GCXP$\}$ and $Y \in \{P, O\}$.

(3) DT-X_Y is XP parameterized by *xp_size* for all $X \in \{$ LAXP, GAXP, GCXP$\}$ and $Y \in \{P, O\}$.

However, the techniques employed to obtain our main results (1) and (2) are very different from the techniques employed for the deterministic setting. In particular, (Ordyniak et al. 2024) used an algorithmic meta theorem that does not cover the counting required in the probabilistic setting. We provide a constructive and unified algorithm for all considered

problems that additionally largely improves upon the runtimes obtained in (Ordyniak et al. 2024) for the deterministic setting.

5 Algorithmic Results

Here we show our algorithmic results. That is, we show that DT-X_Y is fixed-parameter tractable parameterized by *mnl_size* and that DS-X_Y and DL-X_Y are fixed-parameter tractable parameterized by *terms_elem* for every $X \in \{\text{LAXP, GAXP, GCXP}\}$ and $Y \in \{P, O\}$. Indeed, we provide a general framework that provides a unified algorithm for all the considered models and explanation types in the probabilistic and outlier setting. To achieve this, we develop the concept of a featuristic function $H_T(e)$ that captures the relevant properties of the different model types in a unified manner. Informally, the featuristic function provides a compact representation of the model tailor-made for answering explanation queries and we show that given a model M of any of the considered types $\mathcal{M} \in \{\text{DT}, \text{DL}, \text{DS}\}$ and a classification c, we can efficiently compute a featuristic function $H_T(e)$ that for every partial example e returns the number of examples e' that agree with e such that $M(e') = c$.

By analyzing the properties of the featuristic function (in Lemmas 5.7 and 5.8), we reduce the number of partial examples required to compute the maximum value of H_T over all partial examples of a given size. To achieve this, we use the fact that H_T is defined on a small set T of terms and features that occur in the same manner in all those terms can be treated almost uniformly for the analysis of H_T. This allows us to analyse only a bounded number of feature types, i.e., two features have the same type if they occur in the same manner in all terms of T. We then show that we only need to consider partial examples that behave almost uniformly with respect to almost all feature types. Finally, in Theorems 5.11 and 5.12 we carefully define the maximization problems for H_T that allow us to solve all of the considered explanation types both in the probabilistic and in the outlier setting.

We start by introducing some required notation as well as featuristic functions. Let t be a term over a set F of features and let e be a partial example over F. We denote by $A(t, e)$ the number of all examples over F that agree with e and t. Then, $A(t, e)$ is given by setting $A(t, e) = 0$ if t and e are contradictory and otherwise $A(t, e) = 2^{|F|-|F(e)\cup F(t)|}$. We start with some simple observations.

Observation 5.1. *Let $f \in F$, $d \in \{0, 1\}$, t be a term, e be a partial example over $F \setminus \{f\}$, and let e^+ be the partial example obtained from e after assigning f to d. Then: $A(t, e^+) = A(t, e)/2$, $A(t \cup \{(f, d)\}, e^+) = A(t \cup \{(f, d)\}, e)$, and $A(t \cup \{(f, 1 - d)\}, e^+) = 0$.*

For the remainder of this section let T be a set of terms over a set F of features. We denote by $\mathcal{P}(T)$ the set of all non-contradictory terms obtained from the union of any subset of T, i.e., $\mathcal{P}(T) = \{ \bigcup T' \mid T' \subseteq T \wedge |\bigcup T'| = |F(\bigcup T')| \}$. Note that $\emptyset \in \mathcal{P}(T)$ and $|\mathcal{P}(T)| \leq 2^{|T|}$.

We say that two features f_1 and f_2 in F are *equivalent w.r.t. T*, denoted by $f_1 \sim_T f_2$, if for every term $t \in T$, it holds that t contains a literal (f_1, b) if and only if t contains the literal (f_2, b) for every $b \in \{0, 1\}$. Then, $\sim_T$ is an equivalence relation on the set F and we denote by $\mathcal{F}_{\sim_T}$ the set of equivalence classes of $\sim_T$. For a partial exampe e and two distinct features f and f', we denote by $S_{f,f'}(e)$ the partial example obtained from e after switching the assignments of f and f', i.e., the example assigns f to $e(f')$ and f' to $e(f)$. Informally, the following lemma shows that features of the same type behave in the same manner for our purposes.

Lemma 5.2 ($\star$). *Let T be a set of terms over F, let f and f' be two features with $f \sim_T f'$, and let e be a partial example. Then, $A(\bigcup T', e) = A(\bigcup T', S_{f,f'}(e))$ for every $T' \subseteq T$.*

We say that $H_T : E_P(F) \to \mathbb{Z}$, where $E_P(F)$ is the set of all partial examples over F, is a *featuristic* function if it is defined as a weighted sum:

$$H_T(e) = \sum_{T' \subseteq T} h_{T'} A(\bigcup T', e),$$

where $h_{T'} \in \mathbb{Z}$ and $h_{T'}$ can be represented by at most $\mathcal{O}(|F|)$ bits for every $T' \subseteq T$ and serves as the coefficient associated with $A(\bigcup T', e)$.

The following lemma now shows that we can compactly represent a model M of any of the considered types together with a classification c in terms of a featuristic function $H_T(e)$ that for every partial example e returns the number of examples e' that agree with e such that $M(e') = c$.

Lemma 5.3 ($\star$). *Let $\mathcal{M} \in \{\text{DT}, \text{DS}, \text{DL}\}$, let $c \in \{0, 1\}$, and let M be a $\mathcal{M}$ over a set F of features. Then, there is a set of terms T over F and a set $\{ h_{T'} \in \mathbb{Z} \mid T' \subseteq T \}$ of coefficients such that the function $H_T(e) = \sum_{T' \subseteq T} h_{T'} A(\bigcup T', e)$ returns the number of examples over F that agree with e and are classified by M as c. Moreover, $|T| \leq \mathsf{MNL}(M)$ and $|\mathcal{P}(T)| \leq \mathsf{MNL}(M)$ if $\mathcal{M} = \text{DT}$ and otherwise $|T| \leq s$, where s is the number of terms or rules in the DS or DL, respectively. Finally, H_T, i.e., the set T together with the set $\{ h_{T'} \mid T' \subseteq T \wedge h_{T'} \neq 0 \}$ of non-zero coefficients, can be computed in time $\mathcal{O}(\mathsf{MNL}(M))$ if $\mathcal{M} = \text{DT}$ and in time $\mathcal{O}(2^s)$ otherwise.*

Proof Sketch. We start by showing the lemma for the case that $\mathcal{M} = \text{DT}$. Therefore, let $\mathcal{D} = (D, \lambda)$ be a DT over F and without loss of generality assume that $\mathsf{MNL}(\mathcal{D}) = |L_0(\mathcal{D})|$. We start by setting $T = \{ \alpha^l_{\mathcal{D}} \mid l \in L_0(\mathcal{D}) \}$ and then we distinguish two cases dependending on c.

If $c = 0$, then for every $T' \subseteq T$, we set $h_{T'} = 1$ if $|T'| = 1$ and $h_{T'} = 0$ otherwise. Then, $H_T(e) = \sum_{t \in T} A(t, e)$ and we claim that $H_T(e)$ satisfies the claim of the lemma. We first show that $H_T(e)$ is equal to the number of examples e' that agree with e and satisfy $\mathcal{D}(e') = c = 0$. First note that for every $l \in L_0(\mathcal{D})$, $A(\alpha^l_{\mathcal{D}}, e)$ is equal to the number of examples that agree with e and end up in the leaf l of D. Moreover, since $\mathcal{D}$ is a DT every example ends up in exactly one leaf of $\mathcal{D}$ and therefore $\sum_{l \in L_0(\mathcal{D})} A(\alpha^l_{\mathcal{D}}, e) = H_T(e)$ is equal to the number of examples that agree with e and end up in some leaf in $L_0(\mathcal{D})$. Since $L_0(\mathcal{D})$ are the only leaves with label 0 in $\mathcal{D}$, this completes the proof of the claim.

If $c = 1$, then for every $T' \subseteq T$, we set $h_{T'} = 1$ if $T' = \emptyset$, $h_t = -1$ if $|T'| = 1$ and $h_{T'} = 0$ otherwise. Then, $H_T(e) = A(\emptyset, e) - \sum_{t \in T} A(t, e)$ and we claim that $H_T(e)$

satisfies the claim of the lemma, the proof of which can be found in the appendix.

Next, we show the lemma for the case that $\mathcal{M} = \text{DS}$. Therefore, let $\mathcal{S} = (S, b)$ be a DS over F and without loss of generality assume that $b = 0$. We set $T = S$, moreover we again distinguish two cases depending on c.

First consider the case that $c = 1 \neq b$. Then, for every $T' \subseteq T$, we set $h_{T'} = 0$ if $T' = \emptyset$, $h_{T'} = 1$ if $|T'|$ is odd, and $h_{T'} = -1$ if $|T'|$ is even. Then, $H_T(e) = \sum_{\emptyset \neq T' \subseteq T} (-1)^{|T'|+1} A(\bigcup T', e)$. Let $S(t, e)$ denote the set of all examples that agree with t and e; note that $|S(t, e)| = A(t, e)$. Then, $H_T(e) = \sum_{\emptyset \neq T' \subseteq T} (-1)^{|T'|+1} |S(\bigcup T', e)|$ and therefore we obtain from the inclusion-exclusion principle that $H_T(e) = |\bigcup_{t \in T} S(t, e)|$. Therefore, $H_T(e)$ is equal to the number of all examples that agree with e and also agree with some term in $T = S$, which show that $H_T(e)$ calculates the correct thing. The case that $c = 0 = b$ is analogously and can be found in the appendix.

Finally, the proof for DLs is largely based on the function $H^{\text{DS}}_T(e)$ that we constructed above for a DS $(S, 0)$ in the case that $c = 1$. That is, $H^{\text{DS}}_S(e)$ is equal to the number of examples over F that agree with e and also agree with at least one term in S. The crucial observation is then that $H_T(e)$ for a DL $(r_1 = (t_1, c_1), \ldots, r_\ell = (t_\ell, c_\ell))$ can be given as $H_T(e) = \sum_{i \in [\ell] \wedge c_i = c} (H^{\text{DS}}_{T_{\leq i}}(e) - H^{\text{DS}}_{T_{\leq i-1}}(e))$, where $T_{\leq i} = \{t_1, \ldots, t_i\}$. □

Let e be a partial example over F. For the remainder of this section let $\mathcal{F}$ be any refinement of $\mathcal{F}_{\sim_T}$. Moreover, let $\psi^e_T : (\mathcal{F} \times \{0, 1\}) \to [|F|]_0$ be the function that for each $F' \in \mathcal{F}$ and $d \in \{0, 1\}$) returns the number of features in F' that are assigned to d by e, where $[n] = \{1, 2, \ldots, n\}$ and $[n]_0 = [n] \cup \{0\}$ for a natural number n. Note that ψ^e_T provides a compact representation of a partial example that is sufficient for our purposes since it preserves the value of H_T, which is shown by the following lemma.

Lemma 5.4 ($\star$)**.** *Let T be a set of terms over a set of features F, let $H_T : E_P(F) \to \mathbb{Z}$ be a featuristic function, and let $\mathcal{F}$ be any refinement of $\mathcal{F}_{\sim_T}$. Then, for every two partial examples e and e' over F with $\psi^e_{\mathcal{F}} = \psi^{e'}_{\mathcal{F}}$, it holds that $H_T(e) = H_T(e')$.*

We say that a function $\psi : \mathcal{F} \times \{0, 1\} \to [|F|]_0$ is *realizable (via an example e)* if there is a partial example e over F such that $\psi = \psi^e_{\mathcal{F}}$. The following observation provides conditions that allow us to assume realizability.

Observation 5.5. *Let T be a set of terms over a set F of features, let $\mathcal{F}$ be any refinement of $\mathcal{F}_{\sim_T}$, let e be a partial example over F, and let $\psi : (\mathcal{F} \times \{0, 1\}) \to [|F|]_0$ such that $\psi \leq \psi^e_{\mathcal{F}}$. Then, ψ is realizable via an example $e' \subseteq e$.*

Let e be an example over F. We denote by $\Psi(\mathcal{F}, e)$ the set of all functions $\psi \in (\mathcal{F} \times \{0, 1\}) \to [|F|]_0$ such that $\psi(F', d) \in \{0, h^e(F', d)\}$ for every $F' \in \mathcal{F}$ and $d \in \{0, 1\}$, where $h^e(F', d) = |\{f : f \in F' \wedge e(f) = d\}|$. Informally, the functions in $\Psi(\mathcal{F}, e)$ represent partial examples that assign all features of the same type in the same manner. Note that $|\Psi(\mathcal{F}, e)| \leq 4^{|\mathcal{F}|}$ and for each $\psi \in \Psi(\mathcal{F}, e)$, $\psi \leq \psi^e_{\mathcal{F}}$.

Let ψ and ψ' be two functions from $(\mathcal{F} \times \{0, 1\})$ to $[|F|]_0$. We say that $\psi_1 \leq \psi_2$ if $\psi(F', d) \leq \psi'(F', d)$ for every $F' \in \mathcal{F}$ and $d \in \{0, 1\}$. We say that $\psi \leq^1 \psi'$ if $\psi \leq \psi'$ and additionally there is at most one pair (F', d') with $F' \in \mathcal{F}$ and $d \in \{0, 1\}$ such that $\psi(F', d) \neq \psi'(F', d)$. Let $\mathsf{PAIRS}(\Psi)$ be the set of all pairs (ψ, ψ') such that $\psi, \psi' \in \Psi$ and $\psi \leq^1 \psi'$ for a set Ψ of functions $(\mathcal{F} \times \{0, 1\})$ to $[|F|]_0$. As we will see later PAIRS allows us to model partial examples that behave non-uniformly with respect to feature types for at most one such type. Note that $|\mathsf{PAIRS}(\Psi(\mathcal{F}, e))| \leq 2|\mathcal{F}| \cdot 4^{|\mathcal{F}|}$.

Informally, the following lemma shows that if we want to find an optimal local explanation for a given example e of a certain size (given by $|e'|$), then it sufficies to only consider examples $e_\star$ that are *incomplete* for at most one feature type assignment $(F', d) \in \mathcal{F} \times \{0, 1\}$, i.e., $\psi^{e_\star}_{\mathcal{F}}(F', d) \notin \{0, \psi^e_{\mathcal{F}}(F', d)\}$. It is important to note that the lemma only applies for the case that the thought-for partial example does not contradict any term in T, but it provides a crucial step towards showing the general case in Lemma 5.7.

Lemma 5.6. *Let T be a set of terms over a set F of features, let $\mathcal{F}$ be any refinement of $\mathcal{F}_{\sim_T}$, let e be an example over F, let H_T be a featuristic function, and let $e' \subseteq e$ such that e' does not contradict any term in T. Then, there is a pair $(\psi, \psi') \in \mathsf{PAIRS}(\Psi(\mathcal{F}, e))$ and a partial example $e'' \subseteq e$ of size $|e'|$ that is not contradictory with any term in T such that: $\psi \leq \psi^{e''}_{\mathcal{F}} \leq \psi'$ and $H_T(e') \leq H_T(e'')$.*

Proof. We say that a pair $(F', d) \in \mathcal{F} \times \{0, 1\}$ is *bad for $\psi^{e'}_{\mathcal{F}}$* if $\psi^{e'}_{\mathcal{F}}(F', d) \notin \{0, h^e(F', d)\}$. First note that if there is at most one bad pair (F', d) for $\psi^{e'}_{\mathcal{F}}$, then the pair $(\psi, \psi') \in \mathsf{PAIRS}(\Psi(\mathcal{F}, e))$, where ψ (ψ') is obtained from $\psi^{e'}_{\mathcal{F}}$ after changing the value for (F', d) to 0 ($h^e(F', d)$), together with the example $e'' = e'$ satisfy the claim of the lemma. Therefore, our strategy will be to iteratively change e' such as to decrease the number of bad pairs to one.

Intuitively, we construct two partial examples, e_1 and e_2, from e' by modifying the value for two bad pairs. We then show that at least one of these examples, e_1 or e_2, is not worse than e', i.e., $H_T(e') \leq H_T(e_1)$ or $H_T(e') \leq H_T(e_2)$, allowing us to replace e' with the better alternative. Moreover, if $H_T(e') \leq H_T(e_1)$, then it holds that $H_T(e_1) \leq H_T(e^\star_1)$, where $e^\star_1$ is obtained from e_1 in the same manner as e_1 is obtained from e'. Therefore, we can repeat this operation until we remove one of these bad pairs. This process is repeated iteratively until only one bad pair remains.

Let (F_1, d_1) and (F_2, d_2) be two distinct bad pairs for $\psi^{e'}_{\mathcal{F}}$. Then there are $f_1, f'_1 \in F_1$ and $f_2, f'_2 \in F_2$, such that $e'(f_1) = d_1$, $e'(f_2) = d_2$ $f'_1, f'_2 \notin F(e')$, $e(f'_1) = d_1$ and $e(f'_2) = d_2$. Let e_1 be the partial example obtained from e' by replacing the assignment $f_2 = d_2$ with the assignment $f'_1 = d_1$. Note that $e_1 \subseteq e$, because $e' \subseteq e$ and $e(f'_1) = d_1$. Let $\mathcal{T}_1$ be the subset of $\mathcal{P}(T)$ containing all terms that include features from F_1. Because $f_1 \sim_T f'_1$, $e'(f_1) = d_1$, and e' does not contradict any term from T (and therefore from $\mathcal{P}(T)$), we obtain that also e_1 does not contradict any term from $\mathcal{P}(T)$. Similarly, we define e_2 as the partial example obtained from e' by replacing the assignment $f_1 = d_1$ with

the assignment $f'_2 = d_2$ and $\mathcal{T}_2$ to be the subset of $\mathcal{P}(T)$ containing all terms that include features from F_2. Then, also e_2 does not contradict any term from $\mathcal{P}(T)$.

From Observation 5.1 and Lemma 5.2, we obtain that for every $t \in \mathcal{T}_1 \setminus \mathcal{T}_2$, it holds that: $A(t,e) = A(t,e_1)/2 = 2A(t,e_2)$. Intuitively, adding the assignment of features from F_1 or F_2 affects only terms not containing those features, and Observation 5.1 explains how $A(t,e)$ behaves when exchanging equivalent features with respect to T. Similarly, for every $t \in \mathcal{T}'$ with $\mathcal{T}' = (\mathcal{T}_1 \cap \mathcal{T}_2) \cup (\mathcal{P}(T) \setminus (\mathcal{T}_1 \cup \mathcal{T}_2))$, it holds that $A(t,e) = A(t,e_1) = A(t,e_2)$. For $\mathcal{T} \subseteq \mathcal{P}T$, let $H_T^{\mathcal{T}}(e'') = \sum_{T'' \subseteq T \wedge \bigcup T'' \in \mathcal{T}} h_{T''} A(\bigcup T'', e'')$. Then, $H_T(e'') = H_T^{\mathcal{T}'}(e'') + H_T^{\mathcal{T}_1 \setminus \mathcal{T}_2}(e'') + H_T^{\mathcal{T}_2 \setminus \mathcal{T}_1}(e'')$ for any partial example e'' over F. Since $H_T^{\mathcal{T}'}(e') = H_T^{\mathcal{T}'}(e_1) = H_T^{\mathcal{T}'}(e_2)$, $H_T^{\mathcal{T}_1 \setminus \mathcal{T}_2}(e') = \frac{1}{2} H_T^{\mathcal{T}_1 \setminus \mathcal{T}_2}(e_1) = 2H_T^{\mathcal{T}_1 \setminus \mathcal{T}_2}(e_2)$, and $H_T^{\mathcal{T}_2 \setminus \mathcal{T}_1}(e') = 2H_T^{\mathcal{T}_2 \setminus \mathcal{T}_1}(e_1) = \frac{1}{2} H_T^{\mathcal{T}_2 \setminus \mathcal{T}_1}(e_2)$, we obtain that if $H_T^{\mathcal{T}_1 \setminus \mathcal{T}_2}(e') \geq H_T^{\mathcal{T}_2 \setminus \mathcal{T}_1}(e')$, then $H_T(e_1) \geq H_T(e')$ and otherwise $H_T(e_2) \geq H_T(e')$. Moreover, if $H_T(e_i) \geq H_T(e')$, then e_i is a partial example with $|e_i| = |e'|$ that does not contradict any term in $\mathcal{P}(T)$ such that $H_T(e_i) \geq H_T(e')$. Therefore, if we replace e' with e and continue the same process (for the bad pairs (F_1, d_1) and (F_2, d_2)) until one of the bad pairs (F_1, d_1) or (F_2, d_2) disappears, we obtain a partial example $e_\star$, where the number of bad pairs for $\psi_{\mathcal{F}}^{e_\star}$ is strictly smaller than the number of bad pairs for $\psi_{\mathcal{F}}^{e'}$. We can then continue this process for a new pair of bad pairs for $\psi_{\mathcal{F}}^{e_\star}$ until we are left with an example $e'_\star$ that has only one bad pair for $\psi_{\mathcal{F}}^{e'_\star}$. Finally, because $e_1, e_2 \subseteq e$, also $e'_\star \subseteq e$. □

Our next aim is to generalize Lemma 5.6 to find optimal partial examples that can contradict terms in T. Intuitively, the price for this additional generality is that we have to consider examples $e_\star$, where $\psi_{\mathcal{F}}^{e_\star}(F', d)$ is allowed to be 1 for at most $|T|$ feature type assignments $(F', d) \in \mathcal{F} \times \{0,1\}$.

Let e be an example over F. We denote by $\Psi_+(\mathcal{F}, e)$ the set of all functions $\psi \in (\mathcal{F} \times \{0,1\}) \to [|F|]_0$ such that there is $P \subseteq \mathcal{F} \times \{0,1\}$ with $|P| \leq |T|$ satisfying:

- $\psi(F', d) = \min(1, h^e(F', d))$, for every $(F', d) \in P$ and
- $\psi(F', d) \in \{0, h^e(F', d)\}$, for every $(F', d) \in (\mathcal{F} \times \{0,1\}) \setminus P$.

Note that $|\Psi_+(\mathcal{F}, e)| \leq (2|\mathcal{F}|)^{|T|} 4^{|\mathcal{F}|}$ and $|\mathsf{PAIRS}(\Psi_+(\mathcal{F}, e))| \leq 2(2|\mathcal{F}|)^{|T|+1} 4^{|\mathcal{F}|}$, because for every $\psi' \in \Psi_+(\mathcal{F}, e)$, there exists at most $4|\mathcal{F}|$ of $\psi \in \Psi_+(\mathcal{F}, e) \setminus \{\psi'\}$ such that $\psi \leq^1 \psi'$.

The following lemma now already bounds the number of partial examples one has to consider when looking for a local explanation. The main difference to Lemma 5.6 is that we have to deal with terms that are contradicted by the partial example e'. We therefore first analyse the behaviour of e' with respect to terms contradicted by it and then reduce the remaining case to Lemma 5.6.

Lemma 5.7 (⋆). *Let T be a set of terms over a set F of features, let $\mathcal{F}$ be any refinement of $\mathcal{F}_{\sim_T}$, let e be an example over F, let H_T be a featuristic function, and let $e' \subseteq e$. Then there is a pair $(\psi, \psi') \in \mathsf{PAIRS}(\Psi_+(\mathcal{F}, e))$ and a partial example $e'' \subseteq e$ of size $|e'|$ such that: $\psi \leq \psi_{\mathcal{F}}^{e''} \leq \psi'$ and $H_T(e') \leq H_T(e'')$.*

Our next aim is to obtain an equivalent of Lemma 5.7 for global explanations. In particular, the next lemma allows us to bound the number of partial examples that we need to consider for finding an optimal global explanation. In essence, we have to allow at most $|T|$ feature types that are assigned both positively and negatively. However, for those feature types we can assume that there is only one negatively or only one positively assigned feature.

We denote by $\Psi_+(\mathcal{F})$ the set of all functions $\psi \in (\mathcal{F} \times \{0,1\}) \to [|F|]_0$ such that there is an $\mathcal{F}' \subseteq \mathcal{F}$ with $|\mathcal{F}'| \leq |T|$ such that for each $F' \in \mathcal{F}'$, $|F'| > 1$ and $(\psi(F', 0), \psi(F', 1)) \in \{(0,1), (1,0), (1,1), (1, |F'| - 1), (|F'| - 1, 1)\}$; and for each $F' \in \mathcal{F} \setminus \mathcal{F}'$ $(\psi(F', 0), \psi(F', 1)) \in \{(0,0), (0, |F'|), (|F'|, 0)\}$. Note that $|\Psi_+(\mathcal{F})| \leq (5|\mathcal{F}|)^{|T|} 3^{|\mathcal{F}|}$ and $|\mathsf{PAIRS}(\Psi_+(\mathcal{F}))| \leq (5|\mathcal{F}|)^{|T|+1} 3^{|\mathcal{F}|}$, because for every $\psi' \in \Psi_+(\mathcal{F})$, there exists at most $5|\mathcal{F}|$ of $\psi \in \Psi_+(\mathcal{F}) \setminus \{\psi'\}$ such that $\psi \leq^1 \psi'$.

Lemma 5.8 (⋆). *Let T be a set of terms over a set F of features, let $\mathcal{F}$ be any refinement of $\mathcal{F}_{\sim_T}$, let H_T be a featuristic function, and let e' be a partial example over F. Then, there is a pair $(\psi, \psi') \in \mathsf{PAIRS}(\Psi_+(\mathcal{F}))$ and a partial example e'' of size $|e'|$ such that: $\psi \leq \psi_{\mathcal{F}}^{e''} \leq \psi'$ and $H_T(e') \leq H_T(e'')$.*

Previously, we have shown that to find an optimal explanation, it suffices to consider boundely many examples e' bounded by pairs (ψ, ψ') of functions from $\mathsf{PAIRS}(\Psi_+(\mathcal{F}, e))$ (in the case of local explanations) or functions form $\mathsf{PAIRS}(\Psi_+(\mathcal{F}))$ (in the case of global explanations). The following lemma now shows that given such a pair (ψ, ψ'), we can find an optimal explanation bounded by it. We show that this can be achieved rather efficiently having the same time complexity required to compute H_T for one example. Importantly, stating the optimization task using the numbers t_1 and t_2 allow us solve both the probabilitic and the outlier version of the explanation problems. In particular, the probabilitic case is solved by setting $t_1 = 0$ and $t_2 = \gamma$ and the outlier case is solved by setting $t_1 = -\delta$ and $t_2 = 1$.

Lemma 5.9. *Let T be a set of terms over a set F of features, let $\mathcal{F}$ be a refinement of $\mathcal{F}_{\sim_T}$, let H_T be a featuristic function, let t_1 and t_2 be two rationals, and let ψ and ψ' be two realizable functions from $\mathcal{F} \times \{0,1\}$ to $[|F|]_0$ such that $\psi \leq^1 \psi'$. Then, in time $\mathcal{O}(|\mathcal{P}(T)| \cdot |F|)$, we can find a smallest possible partial example e such that $\psi \leq \psi_{\mathcal{F}}^{e} \leq \psi'$ and $H_T(e) \geq t_1 + t_2 2^{|F|-|e|}$, or output correctly that no such example exists.*

Proof. Let $(F', d) \in \mathcal{F} \times \{0,1\}$ be the unique pair such that $\psi(F', d) < \psi'(F', d)$ and let $\Delta = \psi'(F', d) - \psi(F', d)$. Let e_0 be a partial example such that $\psi_{\mathcal{F}}^{e_0} = \psi$, which exists since ψ is realizable. If $H_T(e_0) \geq t_1 + t_2 2^{|F|-|e_0|}$, then we return e_0. Otherwise, for every $i \in [\Delta]$, let e_i be any example obtained from e_{i-1} by assigning a feature from F', which is not yet assigned by e_{i-1}, with value d. Note that $\psi \leq \psi_{\mathcal{F}}^{e_i} \leq \psi'$ and $|e_i| = |e_0| + i$.

From Observation 5.1, Lemma 5.2 and the fact that $\psi_{\mathcal{F}}^{e_1}(F', d) > 0$, we obtain the following for every $i \in [\Delta - 1]$

and $t \in \mathcal{P}(T)$: if $F' \cap F(t) \neq \emptyset$, then $A(t, e_i) = A(t, e_{i+1})$, and otherwise $A(t, e_i) = 2A(t, e_{i+1})$. Let

$$a = \sum_{T' \subseteq T \wedge F' \cap F(\bigcup T') \neq \emptyset} h_{T'} A(\bigcup T', e_\Delta)$$

$$b = \sum_{T' \subseteq T \wedge F' \cap F(\bigcup T') = \emptyset} h_{T'} A(\bigcup T', e_\Delta)$$

then, for every $i \in [\Delta]$, we have that $H_T(e_i) = a + b2^{\Delta - i}$. Note that we can compute a,b and c in $\mathcal{O}(2^{|T|} \cdot |F|)$, because every term $h_{T'} A(\bigcup T', e)$ can be represented by at most $\mathcal{O}(|F|)$ bits.

Then, $H_T(e_i) \geq t_1 + t_2 2^{|F| - |e_i|}$ if and only if $a + b2^{c-i} \geq t_1 + t_2 2^{|F| - |e_0| - i}$ and solving this inequality for i, we obtain $i \geq \lceil \log_2(\frac{t_2 2^{|F|-|e_0|} - b2^c}{(a - t_1)}) \rceil$, which implies that $\lceil \log_2(\frac{t_2 2^{|F|-|e_0|} - b2^c}{(a - t_1)}) \rceil$ is the smallest i that satisfies the original inequality and we can compute this number in $\mathcal{O}(|F|)$. If $i \in [\Delta]$, we return e_i and otherwise we return that no such example exists; this is correct due to Lemma 5.4. □

The next lemma now puts everything together and provides our algorithms to compute partial examples optimizing H_T for local explanations (the first case with given example e) and for global explanations.

Lemma 5.10 ($\star$). *Let T be a set of terms over a set F of features, let H_T be a featuristic function, let t_1 and t_2 be two rationals, and let e be an example.*

In $\mathcal{O}(6^{|T|} \cdot 3^{|T|^2} \cdot 4^{3^{|T|}} \cdot |\mathcal{P}(T)| \cdot |F|)$ time, we can find the smallest partial example $e' \subseteq e$ that satisfies

$$H_T(e') \geq t_1 + t_2 \cdot 2^{|F| - |e'|},$$

or output that such an example does not exist.

Moreover, in $\mathcal{O}(15^{|T|} \cdot 3^{|T|^2} \cdot 2^{3^{|T|}} \cdot |\mathcal{P}(T)| \cdot |F|)$ time, we can find the smallest partial example e' (without the restriction of being a subset of e) that satisfies the above equation or output that such an example does not exist.

Proof Sketch. First we compute $\mathcal{F} = \mathcal{F}_{\sim_T}$, in $\mathcal{O}(|T| \cdot |F|)$ time. Then, for every $(\psi, \psi') \in \mathsf{PAIRS}(\Psi_+(\mathcal{F}, e))$, we apply Lemma 5.9 to potentially obtain a partial example e' that we will keep. If no such partial example is found, we return that no example exists. Otherwise, let e' be a partial example with the smallest size among all those kept. It could be that $e' \not\subseteq e$, but we know that $\psi^{e'}_{\mathcal{F}} \leq \psi^{e}_{\mathcal{F}}$ because of the construction of $\Psi_+(\mathcal{F}, e)$. Therefore, we use Observation 5.5 to return the solution.

Note that $|\mathcal{F}| \leq 3^{|T|}$, and we have at most $2(2|\mathcal{F}|)^{|T|+1} \cdot 2^{2|\mathcal{F}|}$ pairs to check from $\mathsf{PAIRS}(\Psi_+(\mathcal{F}, e))$. Checking one pair requires $\mathcal{O}(|\mathcal{P}(T)| \cdot |F|)$ time. Thus, the complexity of this algorithm is $\mathcal{O}((2|\mathcal{F}|)^{|T|+1} \cdot 2^{2|\mathcal{F}|} \cdot |\mathcal{P}(T)| \cdot |F| + |T| \cdot |F|) = \mathcal{O}(6^{|T|} \cdot 3^{|T|^2} \cdot 4^{3^{|T|}} \cdot |\mathcal{P}(T)| \cdot |F|)$. The correctness proof together with the proof for global explanations can be found in the appendix. □

Finally, we obtain our algorithmic results for all considered models and explanation types in the following two theorem that make use of Lemmas 5.3 and 5.10.

Theorem 5.11. *Let $X \in \{\text{LAXP, GAXP, GCXP}\}$ and $Y \in \{P, O\}$. Then,* DT-X_Y *is fixed-parameter tractable parameterized by* mnl_size.

Proof. Let $\mathcal{A}$ be the given DT defined over a set of features F and let $k = \mathsf{MNL}(\mathcal{A})$. Moreover, if $Y = P$, then let γ be the given threshold and set $t_1 = 0$ and $t_2 = \gamma$. Otherwise, let δ be the given bound on the number of outliers, and set $t_1 = -\delta$ and $t_2 = 1$.

If $X =$LAXP, let e be the given example and let $c' = \mathcal{A}(e)$, otherwise let c be the given classification and set $c' = c$ if $X =$GAXP, and $c' = 1 - c$ if $X =$GCXP. Let T be the set of terms and let H_T be the featuristic function obtained from Lemma 5.3 for $\mathcal{A}$ and $c = c'$ in time $\mathcal{O}(k)$.

Note that $|T|, |\mathcal{P}(T)| \leq k$. If $X =$LAXP, then we return the partial example e' that is obtained from Lemma 5.10 for t_1 and t_2 as defined above and e; note that e' always exists since $e' = e$ satisfies the conditions of Lemma 5.10. Moreover, this step takes time $\mathcal{O}(6^{|T|} \cdot 3^{|T|^2} \cdot 4^{3^{|T|}} \cdot |\mathcal{P}(T)| \cdot |F|) = \mathcal{O}(6^k \cdot 3^{k^2} \cdot 4^{3^k} \cdot k \cdot |F|)$, which also equals the overall complexity for this case.

Otherwise, i.e., if $X \in \{\text{GAXP, GCXP}\}$, then we return the partial example e' that is obtained from Lemma 5.10 (without restriction over e) for t_1 and t_2 as defined above; note that e' exists as long as there is an example that is classified c' by $\mathcal{A}$. Moreover, this step takes time $\mathcal{O}(15^{|T|} \cdot 3^{|T|^2} \cdot 2^{3^{|T|}} \cdot |\mathcal{P}(T)| \cdot |F|) = \mathcal{O}(15^k \cdot 3^{k^2} \cdot 2^{3^k} \cdot k \cdot |F|)$, which also equals to overall complexity for this case. □

The proof of the following theorem now follows along very similar lines as the proof of Theorem 5.11.

Theorem 5.12 ($\star$). *Let $\mathcal{M} \in \{\text{DS}, \text{DL}\}$, let $X \in \{\text{LAXP, GAXP, GCXP}\}$, and $Y \in \{P, O\}$. Then, $\mathcal{M}$-X_Y is fixed-parameter tractable parameterized by* terms_elem.

Finally, we show that all considered problems are solvable in polynomial-time for DTs if the size of the explanation is constant.

Theorem 5.13 ($\star$). *Let $X \in \{\text{LAXP, GAXP, GCXP}\}$ and $Y \in \{P, O\}$. Then,* DT-X_Y *is in* XP *parameterized by* xp_size.

6 Conclusion

We have pioneered the study of the parameterized complexity of computing probabilistic explanations and obtained a comprehensive picture for the fundamental models of DTs, DSs, and DLs. Our unified algorithmic framework, based on featuristic functions, provides efficient methods for computing both local and global probabilistic explanations when parameterized by natural structural properties.

While our results provide a complete picture for DTs, DSs, and DLs, we believe that the study of the probabilistic setting is still in its infancy and there remain many promising directions for future work. In particular, can the algorithmic results obtained in Ordyniak et al. (2024) for (Ordered) Binary Decision Diagrams (OBDD) as well as ensembles of DTs, DSs, DLs, and OBDDs be lifted to the probabilistic setting? Is it possible to furher improve the runtime of our algorithms to, e.g., single-exponential time?

Acknowledgments

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

References

Barceló, P.; Monet, M.; Pérez, J.; and Subercaseaux, B. 2020. Model Interpretability through the lens of Computational Complexity. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dabrowski, K. K.; Eiben, E.; Ordyniak, S.; Paesani, G.; and Szeider, S. 2024. Learning Small Decision Trees for Data of Low Rank-Width. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada*, 10476–10483. AAAI Press.

Darwiche, A.; and Ji, C. 2022. On the Computation of Necessary and Sufficient Explanations. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual Event, February 22 - March 1, 2022*, 5582–5591. AAAI Press.

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

Eiben, E.; Ordyniak, S.; Paesani, G.; and Szeider, S. 2023. Learning Small Decision Trees with Large Domain. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, 3184–3192. ijcai.org.

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.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-Based Explanations for Machine Learning Models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 1511–1519. AAAI Press.

Ignatiev, A.; Narodytska, N.; NicholasAsher; and Marques-Silva, J. 2020. From Contrastive to Abductive Explanations and Back Again. *Proc. AIxIA 2020*, 12414: 335–355.

Izza, Y.; Huang, X.; Ignatiev, A.; Narodytska, N.; Cooper, M. C.; and Marques-Silva, J. 2023. On computing probabilistic abductive explanations. *Int. J. Approx. Reason.*, 159: 108939.

Marques-Silva, J. 2023. Logic-Based Explainability in Machine Learning. *Reasoning Web. Causality, Explanations and Declarative Knowledge*, 24–104.

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

Ordyniak, S.; Paesani, G.; Rychlicki, M.; and Szeider, S. 2024. A General Theoretical Framework for Learning Smallest Interpretable Models. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada*, 10662–10669. AAAI Press.

Ordyniak, S.; Paesani, G.; and Szeider, S. 2023. The Parameterized Complexity of Finding Concise Local Explanations. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, 3312–3320. ijcai.org.

Ordyniak, S.; and Szeider, S. 2021. Parameterized Complexity of Small Decision Tree Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*, 6454–6462. AAAI Press.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Krishnapuram, B.; Shah, M.; Smola, A. J.; Aggarwal, C. C.; Shen, D.; and Rastogi, R., eds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144. ACM.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A Symbolic Approach to Explaining Bayesian Network Classifiers. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 5103–5111. ijcai.org.