Fixed-Parameter Tractability of Learning Small Decision Trees (full paper)

Anonymous author
Anonymous affiliation

Abstract

We consider the NP-hard problem of finding a smallest decision tree which represents a given partially defined Boolean formula. We establish fixed-parameter tractability of the problem with respect to the NLC-width of the instance. We formulate a dynamic programming procedure which utilizes the NLC-decomposition of the instance. For this to work, we establish a succinct representation of partial solutions, so that the space and time requirements of each dynamic programming step remain bounded in terms of the NLC-width.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms → Parameterized complexity and exact algorithms → Fixed parameter tractability

Keywords and phrases parameterized complexity, NLC-width, rank-width, decision trees, partially defined Boolean formulas

1 Introduction

Decision trees have proved to be extremely useful tools for the describing, classifying, generalizing data [18, 22, 25]. In this paper, we consider decision trees for *classification instances (CIs)*, consisting of a finite set E of *examples* (also called *feature vectors*) over a finite set F of *features*. Each example $e \in E$ is a function $e : F \to \{0, 1\}$ which determines whether the feature f is true or false for e. Moreover, E is given as a partition $E^+ \uplus E^-$ into positive and negative examples. For instance, examples could represent medical patients and features diagnostic tests; a patient is positive or negative corresponding to whether they have been diagnosed with a certain disease or not. CIs are also called *partially* or *incompletely defined Boolean functions*, as we can consider the features as Boolean variables, and examples as truth assignments that evaluate to 0 (for positive examples) or 1 (for negative examples). CIs have been studied as a key concept for the logical analysis of data and in switching theory [4, 6, 5, 7, 8, 17, 20].

Because of their simplicity, decision trees are particularly attractive for providing interpretable models of the underlying CI, an aspect whose importance has been strongly emphasized over the recent years [10, 12, 15, 19, 21]. In this context, one prefers *small trees*, as they are easier to interpret and require fewer tests to make a classification. Small trees are also preferred in view of the parsimony principle (Occam's Razor) since small trees are expected to generalize better to new data [2]. However, finding a small decision tree, as formulated in the following decision problem, is NP-complete [16].

Minimum Decision Tree Size (DTS): given a CI $E = E^+ \uplus E^-$ and an integer s, is there a decision tree with at most s nodes for E?

Given this complexity barrier, we propose a fixed-parameter algorithm for the problem, which exploits the input CI's hidden structure. The *incidence graph* of a CI is the bipartite graph $G_I(E)$ whose vertices are the examples on one side and the features on the other, where an example e is adjacent with a feature f if and only if $e(f) = 1$. Figure 1 shows a CI and a smallest decision tree for it, as well as the incidence graph.

Key to our algorithm are new notions for succinctly representing decision trees that correspond to subtrees of the incidence graph's tree decomposition. Based on that, we can carry out a dynamic programming (DP) procedure along the tree decomposition.

While the DP approach using treewidth is quite well understood and can often be quite easily designed for problems on graphs (or more generally problems whose solutions can be represented in terms of the graph for which the tree decomposition is given), the same DP approach can become rather involved if applied to problems whose solutions have no or only minor resemblence to the graph for which one is given a tree decomposition. Probably the most prominent example for this is the celebrated result by Bodlaender [3], where he uses a

E	f_1	f_2	f_3	f_4
$e_1 \in E^-$	0	0	1	0
$e_2 \in E^-$	0	0	1	1
$e_3 \in E^-$	0	1	1	0
$e_4 \in E^-$	1	1	0	0
$e_5 \in E^+$	1	0	0	1
$e_6 \in E^+$	1	0	1	1

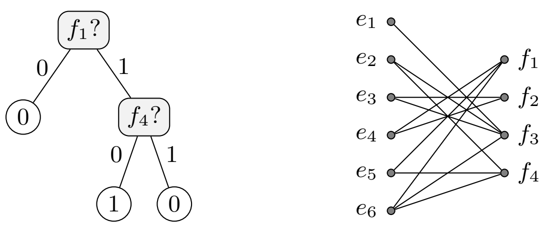


Figure 1 A CI $E = E^+ \uplus E^-$ with six examples and four features (left), a decision tree with 5 nodes that classifies E (middle), the incidence graph $G_I(E)$ (right).

DP approach on an approximate tree decomposition to compute the exact treewidth of a graph; here, the solutions are tree decompositions, which are complex structures that cannot easily be represented in terms of the graph. Other prominent examples include a DP approach to compute the exact treedepth [26] or clique-width [14] using an optimal tree decomposition. We face a similar problem, since solutions in our case are decision trees that do not bear any resemblence to the incidence graph for which we are given the tree decomposition. The main obstacle to overcome, therefore, is the design of the DP-records for our DP algorithm. That is, a record for a node b in a tree decomposition for the incidence graph of E needs to provide a compact representation of partial solutions, i.e. partial solutions in the sense that they represent the part of the solution for the whole instance E that corresponds to the sub-instance induced by all features and examples contained in the bags in the subtree of the tree decomposition rooted at the current node b. We overcome this obstacle in Section 3, where we also provide intuitive descriptions and motivation for the definition of the records (Subsection 3.1).

2 Preliminaries

2.1 Parameterized Complexity

We give some basic definitions of Parameterized Complexity and refer for a more in-depth treatment to other sources [9, 13]. Parameterized complexity considers problems in a two-dimensional setting, where a problem instance is a pair (I, k), where I is the main part and k is the parameter. A parameterized problem is *fixed-parameter tractable* if there exists a computable function f such that instances (I, k) can be solved in time $f(k)\|I\|^{O(1)}$.

2.2 Graphs and NLC-width

We will assume that the reader is familiar with basic graph theory (see, e.g. [11, 1]). We consider (vertex and edge labelled) undirected graphs. Let $G = (V, E)$ be an undirected graph. We write $V(G) = V$ and $E(G) = E$ for the sets of vertices and edges of G, respectively. We denote an edge between $u \in V$ and $v \in V$ as $\{u, v\}$. For a set $V' \subseteq V$ of vertices we let $G[V']$ denote the graph induced by the vertices in V', i.e. $G[V']$ has vertex set V' and edge set $E \cap \{\, \{u, v\} \mid u, v \in V' \,\}$ and we let $G - V'$ denote the graph $G[V \setminus V']$. For a set $E' \subseteq E$ of edges we let denote $G - E'$ the graph with vertex set V and edge set $E \setminus E'$.

A *k-graph* is a pair (G, λ), where $G = (V, E)$ is an undirected graph and $\lambda : V \to [k]$ is a *vertex label mapping* that labels every vertex $v \in V$ with a label $\lambda(v)$ from $[k]$. We call the k-graph consisting of exactly one vertex v (say, labeled by i) an *initial k-graph* and denote it by $i(v)$.

Node label control-width (*NLC-width*) is a graph parameter, defined as follows [28]: Let $k \in \mathbb{N}$ be a positive integer. An *k-NLC-expression tree* of a graph $G = (V, E)$ is a subcubic tree B, where every node b of B is associated with a k-graph (denoted by (G_b, λ_b)), such that:

1. Every leaf represents an initial k-graph $i(v)$ with $i \in [k]$ and $v \in V$.
2. Every non-leaf node b with one child c is a *relabeling node* and is associated with a relabeling function $R_b : [k] \to [k]$. Moreover, G_b is obtained from G_c after relabelling all vertices of G_c with label i to label $R_b(i)$ for every $i \in [k]$.
3. Every non-leaf node b with two children, i.e., a left child l and a right child r, is a *join node* and is associated with a *join matrix*, i.e., a binary $k \times k$ matrix M_b. Moreover,

(G_b, λ_b) is obtained from the disjoint union of (G_l, λ_l) and (G_r, λ_r) after adding an edge from all vertices labeled i in G_l to all vertices labeled j in G_r whenever $M_b[i, j] = 1$.

4. G is equal to the G_r for the root node r of B.

The NLC-width of a graph G, denoted by $nlcw(G)$, is the minimum k for which G has a k-NLC-expression tree. A k-NLC-expression tree is *nice* if every relabelling node has a relabelling function $R : [k] \to [k]$ such that for some $i, j \in [k]$, $R(i) = j$ and $R(\ell) = \ell$ for all $\ell \in [k] \setminus \{i\}$. Clearly, given a k-NLC-expression tree, a nice k-NLC-expression tree can be found in polynomial time; simply replace every relabelling node (that relabels more than one label at a time) by a sequence of relabelling nodes.

Let b be a node in a k-NLC-expression tree of a graph G. We denote by V_b the set of vertices of G_b. By the definition of a k-NLC-expression tree, if $u, v \in V_b$ have the same label in (G_b, λ_b) and $w \in V(G) \setminus V_b$, then u is adjacent to w in G if and only if v is.

Computing the NLC-width of a graph is NP-hard [?]. However, it is sufficient to use the algorithm of Seymour and Oum [?], which returns a c-expression for some $c \leq 2^{3\text{cw}(G)+2} - 1$ in $O(n^9 \log n)$ time, or the later improvements of Oum [24] and Hliněný and Oum [?] that provide cubic-time algorithms which yield a c-expression for some $c \leq 8^{cw(G)} - 1$ and $c \leq 2^{\text{cw}(G)+1} - 1$, respectively.

2.3 Classification Problems

An *example* e is a function $e : feat(e) \to \{0, 1\}$ defined on a finite set $feat(e)$ of *features*. For a set E of examples, we put $feat(E) = \bigcup_{e \in E} feat(e)$. We say that two examples e_1, e_2 *agree* on a feature f if $f \in feat(e_1)$, $f \in feat(e_2)$ and $e_1(f) = e_2(f)$. If $f \in feat(e_1)$, $f \in feat(e_2)$ but $e_1(f) \neq e_2(f)$, we say that the examples *disagree on* f.

A *classification instance (CI)* (also called a *partially defined Boolean function* [17]) $E = E^+ \uplus E^-$ is the disjoint union of two sets of examples, where for all $e_1, e_2 \in E$ we have $feat(e_1) = feat(e_2)$. The examples in E^+ are said to be *positive*; the examples in E^- are said to be *negative*. A set X of examples is *uniform* if $X \subseteq E^+$ or $X \subseteq E^-$; otherwise X is *non-uniform*.

Given a CI E, a subset $F \subseteq feat(E)$ is a *support set* of E if any two examples $e_1 \in E^+$ and $e_2 \in E^-$ disagree in at least one feature of F. Finding a smallest support set, denoted by $\text{MSS}(E)$, for a classification instance E is an NP-hard task [17, Theorem 12.2].

We define the *incidence graph* of E, denoted by $G_I(E)$, as the bipartite graph with partition $(E, feat(E))$ having an edge between an example $e \in E$ and a feature $f \in feat(e)$ if $f(e) = 1$.

2.4 Decision Trees

A *decision tree* (DT) (or *classification tree*) is a rooted tree T with vertex set $V(T)$ and arc set $A(T)$, where each non-leaf node (called a *test*) $v \in V(T)$ is labelled with a feature $feat(v)$, each non-leaf node v has exactly two out-going arcs, a *left arc* and a *right arc*, and each leaf is either a *positive* or a *negative* leaf. We write $feat(T) = \{ v \in V(T) \mid feat(v) \}$.

Consider a CI E and a decision tree T with $feat(T) \subseteq feat(E)$. For each node v of T we define $E_T(v)$ as the set of all examples $e \in E$ such that for each left (right, respectively) arc (u, v) on the unique path from the root of T to v we have $e(feat(v)) = 0$ ($e(feat(v)) = 1$, respectively). T *correctly classifies* an example $e \in E$ if e is a positive (negative) example and $e \in E_T(v)$ for a positive (negative) leaf. We say that T *classifies* E (or simply that T is a DT for E) if T correctly classifies every example $e \in E$. See Figure 1 for an illustration of a CI, its incidence graph, and a DT that classifies E.

The size of T is its number of nodes, i.e. $|V(T)|$. We consider the following problem.

MINIMUM DECISION TREE SIZE (DTS)	
Input:	A classification instance E and an integer s.
Question:	Is there a decision tree of size at most s for E?

We now give some simple auxiliary lemmas that are required by our algorithm.

▶ **Lemma 1.** *Let A be a set of features of size a. Then the number of DTs of size at most s that use only features in A is at most a^{2s+1} and those can be enumerated in $\mathcal{O}(a^{2s+1})$ time.*

Proof. We start by counting the number of trees T with n nodes that can potentially underlie a DT with n nodes. Note that there is one-to-one correspondence between trees T that underlie a DT with n nodes and unlabelled rooted ordered binary trees with n nodes (where ordered refers to an ordering of the at most 2 child nodes). Since it is known that the number of unlabelled rooted ordered binary trees with n nodes is equal to the n-th Catalan number C_n and that those trees can be enumerated in $\mathcal{O}(C_n)$ time [27], we already obtain that we can enumerate all of the at most C_n possible trees T underlying a DT of size n in $\mathcal{O}(C_n)$ time. Therefore, there are at most sC_s possible trees of size at most s that can underlie a DT with at most s nodes and those can be enumerated in $\mathcal{O}(sC_s)$ time. It now remains to bound the number of possible feature assignments $\mathit{feat}(f)$ for these trees as well as the number of possibilities for the leave nodes that can be either labelled positive or negative. Since we can assume that $a \geq 2$, we obtain that the number of possible feature assignments (and labellings of leaf-nodes) of a tree T with n nodes is at most a^n. Taking everything together, we obtain that there are at most $sC_sa^s \leq s4^sa^s \leq a^{2s+1}$ many DTs of size at most s using only features in A and those can be enumerated in $\mathcal{O}(a^{2s+1})$ time. ◀

▶ **Lemma 2.** *Let A be a set of features of size a. There are at most $a^{2^{a+1}+3}$ inclusion-wise minimal DTs using only features in A and these can be enumerated in $\mathcal{O}(a^{2^{a+1}+3})$ time.*

Proof. Note that an inclusion-wise minimal DT T that uses only features in A has at most 2^a+1 nodes; this is because every feature appears at most once on every path T. Therefore, we obtain from Lemma 1 that the number of choices for T is at most $a^{2(2^a+1)+1} = a^{2^{a+1}+3}$. ◀

▶ **Lemma 3.** *Let E be a CI. Then one can decide whether E has a DT and if so output a DT of minimum size for E in time $\mathcal{O}((2^{|E|})^{4|E|-1})$.*

Proof. Note first that $|\mathit{feat}(E)| \leq 2^{|E|}$ since we can assume that E does not contain two equivalent features. Moreover, E has a DT if and only if $\mathit{feat}(E)$ is a support set, which can be checked in time $\mathcal{O}(|E|^2|\mathit{feat}(E)|)$ by checking, for every pair of positive and negative examples in E, whether there is a feature that distinguishes them. If this is not the case, we output **NO**, so assume that E has a DT. Note that any inclusion-wise minimal DT for E has at most $|E|$ leaves and therefore size at most $2|E|-1$. We can therefore employ Lemma 1 to enumerate all inclusion-wise minimal potential DTs for E in time $\mathcal{O}((2^{|E|})^{2(2|E|-1)+1}) \in \mathcal{O}((2^{|E|})^{4|E|-1})$. For every such tree we then check whether it is indeed a DT for E and return a DT for E of minimum size found during this process. ◀

3 An FPT-Algorithm for NLC-width

In this section, we present our main result, i.e. we will show that DTS is fixed-parameter tractable parameterized by NLC-width.

▶ **Theorem 4.** *Let E be a CI, let B be an NLC-decomposition of width ω for $G_I(E)$, and let s be an integer. Then, deciding whether E has a DT of size at most s is fixed-parameter tractable parameterized by ω.*

todo: Due to proposition ...

▶ **Corollary 5.** DTS *is fixed-parameter tractable parameterized by NLC-width.*

In principle, we will use a dynamic programming algorithm along the NLC-decomposition (B, χ) of $G_I(E)$ that computes a set of records for every node b of B in a bottom-up manner. Each record will represent an equivalence class of solutions (DTs) for the whole instance restricted to the examples and features contained in the current subtree rooted in b, i.e. the examples and features contained in $\chi(b)$. Before we continue with the formal notions and definitions required to define the records, we want to illustrate the main ideas and motivations. In what follows let B be an NLC-decomposition of $G_I(E)$ of width k. For $b \in V(B)$, we write $\mathit{feat}(b)$ and $\mathit{exam}(b)$ for the sets $\chi(b) \cap \mathit{feat}(E)$ and $\chi(b) \cap E$, respectively.

3.1 Description of the Main Ideas Behind the Algorithm

todo: adjust to NLC-width

Consider a node b of B. To simplify the presentation, we will sometime refer to the features and examples in $\chi(B_b) \setminus \chi(b)$ as *forgotten* features and examples and we refer to the features and examples in $(\mathit{feat}(E) \cup E) \setminus \chi(B_b)$ as *future* features and examples. We start with some simple observations that follow immediately from the properties of tree decompositions.

▶ **Observation 6.***(1)* $e(f) = 0$ *for every forgotten example* $e \in \mathit{exam}(B_b) \setminus \mathit{exam}(b)$ *and future feature* $f \in \mathit{feat}(E) \setminus \mathit{feat}(B_b)$,

(2) $e(f) = 0$ *for every future example* $e \in E \setminus \mathit{exam}(B_b)$ *and forgotten feature* $f \in \mathit{feat}(B_b) \setminus \mathit{feat}(b)$;

Proof. Towards showing (1), let e be an example in $\mathit{exam}(B_b) \setminus \mathit{exam}(b)$ and let f be a feature in $\mathit{feat}(E) \setminus \mathit{feat}(B_b)$. We claim that because (T, χ) is a tree decomposition of $G_I(E)$, the graph $G_I(E)$ cannot contain an edge between e and f, which implies that $e(f) = 0$. Suppose for a contradiction that this is not the case, i.e. $\{e, f\} \in E(G_I(E))$. Then, because of property (T1) of a tree decomposition, there must exist a node b' such that $e, f \in \chi(b')$. But then, if $b' \in V(B_b)$ we obtain that $f \notin \chi(b')$. Similarly, if $b' \in V(B \setminus B_b)$, we obtain that $e \notin \chi(b')$ since otherwise e would violate property (T2) of a tree decomposition. This completes the proof for (1); the proof for (2) is analogous. ◀

Informally, Observation 6 shows that forgotten examples cannot be distinguished by future features and future examples cannot be distinguished by forgotten features. Consider a DT T for E and a node b of B. For a set W containing features and examples from E, we denote by $E[W]$ the sub-instance of E induced by the features and examples in W. Our aim is to obtain a compact representation (represented by records) of the partial solution for the sub-instance $E[\chi(B_b)]$ of E induced by the features and examples in $\chi(B_b)$ represented by T.

Intuitively, such a compact representation has to (1) represent a partial solution (DT) for the examples in $\mathit{exam}(B_b)$ and (2) retain sufficient information about the structure of T in order to decide whether it can be extended to a DT that also classifies the examples in $E \setminus \mathit{exam}(B_b)$.

For illustration purposes let us first consider the simplified case that $\mathit{exam}(b) = \emptyset$. Because of Observation 6 (1), this implies that every forgotten example goes to the left child of any node t in T that is assigned a future feature. Therefore, under the assumption that $\mathit{exam}(b) = \emptyset$ the DT T' obtained from T after:

- removing the subtree T_r of T for every right child r of a node t of T with $\mathit{feat}(t) \in \mathit{feat}(E) \setminus \mathit{feat}(B_b)$ and replacing t with an edge from its parent in T to its left child in T

is a DT for $E[\chi(B_b)]$. Note that this means that under the rather strong assumption that $\mathit{exam}(b) = \emptyset$, the part of T that takes care of the sub-instance $E[\chi(B_b)]$ is itself a DT using only features in $\mathit{feat}(B_b)$; we will see later that unfortunately this is no longer the case if $\mathit{exam}(b) \neq \emptyset$. Note that even though T' is a DT for $E[B_b]$, it does not yet constitute a compact representation, since the number of features it uses in $\mathit{feat}(B_b) \setminus \mathit{feat}(b)$ is potentially unbounded. However, we obtain from Observation 6 (2) that every future example will end up in the left child of every node t of T' that is assigned a forgotten feature. This means that to decide whether T' can be extended to a DT for the whole instance, the nodes that are assigned forgotten features are not important. In fact, the only nodes in T' that can be important for the classification of future examples are the nodes that are assigned features in $\mathit{feat}(b)$. That is, it is sufficient to remember the DT T'' obtained from T' after:

- removing the subtree T_r of T' for every right child r of a node t of T' with $\mathit{feat}(t) \in \mathit{feat}(B_b) \setminus \mathit{feat}(b)$ and replacing t with an edge from its parent in T' to its left child in T'.

Since the number of possible DT T'' is clearly bounded in terms of the number of features in $\mathit{feat}(b)$ (and therefore in terms of the treewidth of $G_I(E)$), this would already give us the compact representation that we are looking for. However, this only works in the case that $\mathit{exam}(b) = \emptyset$, which is clearly not the case in general.

So let us now consider the general case with $\mathit{exam}(b) \neq \emptyset$. The first difference now is that the part of T that takes care of the sub-instance $E[\chi(B_b)]$ is no longer a DT that only uses features in $\mathit{feat}(B_b)$. In fact, it could even be the case that $E[\chi(B_b)]$ does not have a DT, because there could exist examples in $\mathit{exam}(b)$ that can only be distinguished using the features in $\mathit{feat}(E) \setminus \mathit{feat}(B_b)$. This means that we have to allow our partial solution for $E[\chi(B_b)]$ to use future features. Fortunately, we do not need to know which exact future feature is used by our partial solution but it suffices to know that a future feature is used and how it behaves w.r.t. the examples in $\mathit{exam}(b)$; this is because Observation 6 (1) implies that a future feature is used in a partial solution only for the purpose of distinguishing examples in $\mathit{exam}(b)$. Moreover, because every forgotten example ends up in the left child of any node t of T that uses a future feature, we only need to remember the left child for those nodes. Also, we only need to remember occurrences of those nodes (using future features) if at least one example in $\mathit{exam}(b)$ ends up to in the right child of such a node; otherwise the node has no influence on the classification of examples in $\mathit{exam}(B_b)$. Finally, we cannot simply forget nodes that use forgotten features (as we could in the case that $\mathit{exam}(b) = 0$). This is because we need to know exactly where the examples in $\mathit{exam}(b)$ end up at. For instance, if such an example in $\mathit{exam}(b)$ ends up in the right child of a node using a future feature, we need to know that this is the case because this means that the example has to be classified in this place at a later stage of the algorithm. Nevertheless, we do not need to remember all occurrences of nodes using forgotten features, but only those for which there is at least one example in $\mathit{exam}(b)$ that ends up in the right child of the node. Similarly, we do not need to remember the exact forgotten feature that is used but only how it behaves towards the examples in $\mathit{exam}(b)$. In summary, we only need to remember the full information about the nodes of T that use a feature in $\mathit{feat}(b)$. For all other nodes, i.e. nodes that use either forgotten or future features, we only need to remember such a node, if at least one example in $\mathit{exam}(b)$ ends up in its right child. Moreover, even if this is the case, we only need to remember the following for such nodes:

- whether it uses a future or a forgotten feature and

- how it behaves w.r.t. the examples in *exam*(b).

With these ideas in mind, we are now ready to provide a formal definition of the compact representation of the part of T that takes care of the sub-instance $E[\chi(B_b)]$.

3.2 Formal Definition of Records and Preliminary Results

In the following, let E be a CI and let B be a k-NLC-expression tree for $G_I(E)$. Consider a node b of B. Recall that b is either a leaf node associated with a k-graph $i(v)$, a relabelling node with 1 child and with relabelling function R_b, or a join node with a left child, a right child and a join matrix M_b. Moreover, recall that (G_b, λ_b) is the k-graph associated with b (whose unlabeled version is a subgraph of G) and V_b is the set of vertices of G_b. Additionally, we will use the following notation. We denote by $\mathit{feat}(b)$ the set $V_b \cap \mathit{feat}(E)$ of features in V_b and by $\mathit{exam}(b)$ the set $V_b \cap E$ of examples in V_b.

Consider a node b of B. Let L be a set of labels (usually $L = [k]$). For a subset $L' \subseteq L$, we denote by $\overline{L'}$ the set $L \setminus L'$. For a label $l \in L$, we introduce a new feature f_l, which we will call a *forgotten feature*. Moreover, for a subset $L' \subseteq L$ of labels, we introduce a new feature $f_{L'}$, which we call an *future (or introduce) feature*. Let $F_L = \{\, f_l \mid l \in L \,\}$ be the set of all forgotten features and let $I_L = \{\, f_{L'} \mid L' \subseteq L \,\}$ be the set of all future features w.r.t. L. To distinguish features in $\mathit{feat}(E)$ from forgotton and future features, we will refer to them as *real features*.

definition of new features

Let T be a decision tree and $t \in V(T)$. We say that a node t_A is a *left/right anchestor* of t if t is contained in the subtree of T rooted at the left/right child of t_A. We denote by $\mathrm{anc}_L(t)/\mathrm{anc}_R(t)$ the set of all left/right ancestors of t in T. We denote by $\mathrm{anc}(t)$ the set of all *ancestors* of t in T, i.e., $\mathrm{anc}(t) = \mathrm{anc}_L(t) \cup \mathrm{anc}_R(t)$.

Let T be a decision tree and $t \in V(T)$ be an inner node of T with left child l, right child r, and parent p. We say that T' is obtained from T after *left/right-contracting* t if T' is the decision tree obtained from T after removing t together with all nodes in T_r/T_l and adding the edge between p and l/r; if t has no parent then no edge is added.

We say that T is a *decision tree* for b, if T is a decision tree for $\mathit{exam}(b)$ that uses only the features in $\mathit{feat}(b)$. We say that an inner node $t \in V(T)$ is *left/right redundant* in T if $\mathit{feat}(t) \in \mathit{feat}(\mathrm{anc}_L(t))/\mathit{feat}(t) \in \mathit{feat}(\mathrm{anc}_R(t))$. We say that t is redundant if it is either left redundant or right redundant. Intuitively, a node t is left/right redundant if all examples that end up at t, i.e., the examples $E_T(t)$, go the left/right child of t in T. Therefore, if t is left/right redundant in T, then the tree obtained after left/right-contracting t is still a decision tree.

We say that T is a *decision tree template* for b if T is a decision tree for $\mathit{exam}(b)$ that can additionally use the future features in $I_{[k]}$. Here, we assume that a future feature $f_{L'} \in I_{[k]}$ for some $L' \subseteq [k]$ is 1 at an example $e \in \mathit{exam}(b)$ if $\lambda_b(e) \in L'$ and otherwise it is 0. We say that a decision tree template is *complete* if it does not use any features in $I_{[k]}$, otherwise we say that it is *incomplete*. Informally, the role of the future features in a decision tree template is provide spaceholders for the features in $\mathit{feat}(E) \setminus \mathit{feat}(b)$. Because all of those features behave the same w.r.t. to examples in $\mathit{exam}(b)$ having the same label, they can be characterised by the set of labels for which those features are 1. Let T be a decision tree template for b and let $t \in V(T)$. We denote by $A(t)$ the set of *filtered labels* for t, i.e., $A(t) = (\bigcap_{f_{L'} \in \mathit{feat}(\mathrm{anc}_L(t)) \cap I_{[k]}} \overline{L'}) \cap (\bigcap_{f_{L'} \in \mathit{feat}(\mathrm{anc}_R(t)) \cap I_{[k]}} L')$. Informally, $A(t)$ is the set of all labels $l \in [k]$ such that an example e with label l would end up at t, if only the effect of the future features on the path to t is considered. We say that t with $f_{L'} = \mathit{feat}(t) \in I_{[k]}$ is *left/right redundant* in T if $A(t) \subseteq L'/A(t) \subseteq \overline{L'}$. We say that t is *redundant* if it is either left-redundant or right-redundant. Intuitively, t is left/right redundant if all examples that can reach t (considering the influence of the future features only) end up in the left/right child of t. This also implies that if t is left/right redundant then the decision tree obtained after left/right contracting t is equivalent with T (all examples end up in the same leaves).

We say that T is a *decision tree skeleton* for b if T is a decision tree that can only use features in $F_{[k]} \cup I_{[k]}$. Note that because of the features $F_{[k]}$, whose behaviour w.r.t. the examples in $\mathit{exam}(b)$ is not defined, the behaviour w.r.t. the examples in $\mathit{exam}(b)$ of such a DT skeleton is not necessarily defined. Nevertheless, the behaviour of a feature f_l in $F_{[k]}$ is well-defined w.r.t. to the examples in $\mathit{exam}(E) \setminus \mathit{exam}(b)$, i.e., it behaves the same as any feature in $\mathit{feat}(b)$ with label l. Intuitively, decision tree skeletons are obtained from decision tree templates after replacing every feature f in $\mathit{feat}(b)$ with its label $\lambda_b(f)$. This allows us to further compress the information contained in decision tree templates, while still keeping the information about how the decision tree template behaves w.r.t. future examples in $\mathit{exam}(b)$. In particular, decision tree skeletons will form the main information stored by our records.

Let T be a decision tree skeleton and $t \in V(T)$. Similarily as we did for decision tree templates, we say that T is *complete* if it uses no future features and otherwise we say that it is incomplete. We say that an inner node t with $f_l = \mathit{feat}(t) \in F_{[k]}$ is *left/right redundant* in T if $f_l \in \mathit{feat}(\mathrm{anc}_L(t))/f_l \in \mathit{feat}(\mathrm{anc}_R(t))$. Similarily, as for decision tree (templates), if t is left/right redundant, then we can left/right contract t without changing the properties of T.

Let T be a decision tree (skeleton/template). Then, we denote by $r(T)$ the decision tree obtained from T after left/right contracting every left/right redundant node of T. Note that if T is a decision tree (skeleton/template) for b, then so is $r(T)$.

▶ **Observation 7.** *Let T be a decision tree skeleton/template for b. Then, so is $r(T)$.*

a short proof

Proof. ◀

We say that T is *reduced* if $r(T) = T$.

▶ **Lemma 8.** *Let T be a reduced decision tree (skeleton/template) using at most a real features, b forgotten features, and c future features. Then, T has size at most ?.*

todo

Proof. ◀

definition of relabelling

Let T be a decision tree. A *feature relabeling* for T is a function $\alpha : F' \to \mathit{feat}(E) \cup F_L \cup I_L$, where $F' \subseteq \mathit{feat}(T)$ and L is some set of labels (usually $L = [k]$). With a slight abuse of notation, we denote by $\alpha(T)$, the decision tree obtained after relabeling all features in F' (used by T) according to α, i.e., $\alpha(T)$ is obtained from T after replacing the feature assignment function $\mathit{feat}_T(t)$ for T with the function $\mathit{feat}_{\alpha(T)}(t)$ defined by setting $\mathit{feat}_{\alpha(T)}(t) = \alpha(\mathit{feat}_T(t))$ if $\mathit{feat}(t) \in F'$ and $\mathit{feat}_{\alpha(T)}(t) = \mathit{feat}_T(t)$, otherwise. We say that two feature relabellings $\alpha_1 : F_1 \to \mathit{feat}(E) \cup F_L \cup I_L$ and $\alpha_2 : F_2 \to \mathit{feat}(E) \cup F_L \cup I_L$ are *compatible* if they agree on their shared domain $F_1 \cap F_2$.

We denote by α^s_b the *standard feature relabelling* for b, i.e., the function $\alpha^s_b : \mathit{feat}(b) \to [k]$ defined by setting $\alpha^s_b(f) = \lambda_b(f)$ for every $f \in \mathit{feat}(b)$.

Semantics of records

We are now ready to define the records and their semantics. A *record* for b is a pair (T, s) such that T is a reduced decision tree skeleton for b and s is a natural number. We say that a record (T, s) is *valid* for b if s is the minimum number such that there is a (reduced) decision tree template T' for b such that $r(\alpha^s_b(T')) = T$ and $s = |V(T') \setminus V(T)|$. We denote by $\mathcal{R}(b)$ the set of all valid records for b. The following corollary follows immediately from Lemma 8.

▶ **Corollary 9.** $|\mathcal{R}(b)| \leq ?$

Note that E has a DT of size at most s if and only if $\mathcal{R}(r)$ contains a record (T, s) such that T is complete, where r is the root of B. ...

auxilariary properties of feature relabelings and reductions

▶ **Lemma 10.** *Let T be a decision tree and let α be a feature relabelling for T. Then, $r(\alpha(T)) = r(\alpha(r(T)))$.*

▶ **Observation 11.** *Let T be a decision tree and let α_1 and α_2 be two compatible feature relabelling for T. Then, $\alpha_1\alpha_2(T)) = \alpha_2\alpha_1(T))$.*

3.3 Proof to the Main Result

We will now show that we can compute $\mathcal{R}(b)$ for every of the 3 node types of a nice k-NLC expression tree provided that $\mathcal{R}(c)$ has already been computed for every child c of b.

▶ **Lemma 12** (leaf node)**.** *Let $b \in V(B)$ be a leaf node. Then $\mathcal{R}(b)$ can be computed in time* ??.

Proof. Let $i(v)$ be the initial k-graph associated with b. If v is a feature, then $\mathcal{R}(b)$ contains all records $(T, 0)$ such that T is a reduced decision tree skeleton for b using only the features in $\{f_{\lambda(v)}\} \cup I_{[k]}$. The correctness in this case follows because V_b contains no examples and therefore every reduced decision tree skeleton constitutes a valid record for b. Moreover, the run-time follows from Lemma **??**, since the time required to enumerate all those reduced decision tree skeletons is at most $\mathcal{O}(?)$.

If, on the other hand v is an example, then $\mathcal{R}(b)$ contains all records $(T, 0)$ such that T is a reduced decision tree skeleton for b using only the features in $I_{[k]}$ and which correctly classify v. Because of Lemma **??**, those can be enumerated in time $\mathcal{O}(?)$ and checking for each of those whether it correctly classifies v can be achieved in time $\mathcal{O}(?)$.

todo: show correctness

◀

▶ **Lemma 13** (join node)**.** *Let $b \in V(B)$ be a join node. Then $\mathcal{R}(b)$ can be computed in time $\mathcal{O}(k(2k + 2^k + 2)2^{6k+1})$.*

Proof. Let b_L and b_R be the left and right child of b in B, respectively.

Let M_b be the join matrix for the node b, i.e., M_b is a $k \times k$ binary matrix. For every label $i \in [k]$, let $A_{i,*} = \{\, j \in [k] \mid M_b[i, j] = 1 \,\}$ and $A_{*,i} = \{\, j \in [k] \mid M_b[j, i] = 1 \,\}$.

To distiguish between forgotten features from the left and the right subtree, we introduce the left i_L and the right version i_R for every label $i \in [k]$. With a slight abuse of notation, we also denote by $[k_L]$ be the set $\{1_L, \dots, k_L\}$ of (left) labels and we denote by $[k_R]$ be the set $\{1_R, \dots, k_R\}$ of (right) labels.

To compute the set $\mathcal{R}(b)$ of valid record for b, we first enumerate all reduced DT skeletons T using features in $[k_L] \cup [k_R] \cup I_{[k]}$. Because of Lemma 17, those can be enumerated in time $\mathcal{O}((2k + 2^k + 2)2^{3k+1})$.

For every such reduced DT skeleton T, we now do the following in order to decide whether T gives rise to a valid record for b. Let $\alpha^{LR\to} : F_{[k_L]} \cup F_{[k_R]} \to F_{[k]}$ be the feature relabeling that relabels every (left/right) feature $f_{i_H} \in F_{[k_L]} \cup F_{[k_R]}$ (for some $H \in \{L, R\}$) to its original feature f_i.

Let $\alpha^L : F_{[k_R]} \to I_{[k]}$ be the feature relabeling that relabels every forgotten feature $f_{i_R} \in F_{[k_R]}$ to the future feature $f_{A_{*,i}}$. Let T_L be the reduced DT skeleton obtained from T after applying the relabelling using α^L followed by $\alpha^{LR\to}$ and then reducing the resulting DT skeleton, i.e., $T_L = r(\alpha^{LR\to}(\alpha^L(T)))$.

Similarily, let $\alpha^R : F_{[k]_L} \to I_{[k]}$ be the feature relabeling that relabels every forgotten feature $f_{i_L} \in F_{[k_L]}$ to the future feature $f_{A_{i,*}}$. Let T_R be the reduced DT skeleton obtained from T after applying the relabelling using α^R followed by $\alpha^{LR\rightarrow}$ and then reducing the resulting DT skeleton, i.e., $T_R = r(\alpha^{LR\rightarrow}(\alpha^R(T)))$.

Let $\hat{T} = \alpha^{LR\rightarrow}(T)$ and $\hat{s} = |V(T) \setminus V(\hat{T})|$. We now check whether there are records $(T_L, s_L) \in \mathcal{R}(b_L)$ and $(T_R, s_R) \in \mathcal{R}(b_R)$. If not we discard T and if yes, then we add the record $(\hat{T}, s_L + s_R + \hat{s})$ to $\mathcal{R}(b)$. This completes the description about how the records $\mathcal{R}(b)$ are computed. Moreover, the run-time for computing $\mathcal{R}(b)$ can be obtained as follows. First, because of Lemma 17, we can enumerate all reduced DT skeletons T in time $\mathcal{O}((2k + 2^k + 2)2^{3k+1})$. Moreover, computing $\hat{T}$ and $\hat{s}$ can be done in time $\mathcal{O}(|T|) = \mathcal{O}(s)$. Finally, computing T_L and T_R and checking the existence of the records $(T_L, s_L) \in \mathcal{R}(b_L)$ and $(T_R, s_R) \in \mathcal{R}(b_R)$ can be achieved in time $\mathcal{O}(?)$. Therefore, we obtain $\mathcal{O}(?)$ as the total run-time for computing $\mathcal{R}(b)$.

old run-time argument below, should be replaced above

We now show the correctness of our construction for $\mathcal{R}(b)$, i.e., we have to show that a record (T, s) is valid if and only if we have added such a record according to our construction above.

Towards showing the forward direction, suppose that $(\hat{T}, s)$ is a valid record in $\mathcal{R}(b)$. Therefore, there is a DT template T' for b such that $\hat{T} = r(\eta_{\alpha^s_b}(T'))$ and $s = |V(T') \setminus V(T)|$.

Because $\hat{T}$ is obtained from T' by reduction, every node in $\hat{T}$ corresponds to a unique node in T'. Therefore, there is an injective function $z_H : V(\hat{T}) \rightarrow V(T')$ mapping every node in $\hat{T}$ to its original node in T'. Let T be the DT obtained from $\hat{T}$ after by setting $\mathit{feat}_T(t) = i_H$ if $\mathit{feat}_{\hat{T}}(t) = i$ and $\mathit{feat}_{T'}(t) \in \mathit{feat}(b_H)$ for $H \in \{L, B\}$.

Note that $\hat{T} = \eta_{\alpha^{LR\rightarrow}}(T)$ and $\hat{T}$ is reduced because $(\hat{T}, s) \in \mathcal{R}(b)$.

Let $\alpha^{\rightarrow R} : F_{[k]} \rightarrow F_{[k_R]}$ $(\alpha^{\rightarrow L} : F_{[k]} \rightarrow F_{[k_L]})$ be the feature relabeling that relabels every forgotten feature $f_i \in F_{[k]}$ to its corresponding forgotten feature in $[k_R]$ $([k_L])$, i.e., $\alpha^{\rightarrow R}(i) = i_R$ $(\alpha^{\rightarrow L}(i) = i_L)$ for every $i \in [k]$.

Note that $T = r(\eta_{\alpha^{\rightarrow L}}(\eta_{\alpha^s_{b_L}}(\eta_{\alpha^{\rightarrow R}}(\eta_{\alpha^s_{b_R}}(T')))))$.

Let $T_L = r(\eta_{\alpha^L}(T))$ and $T_R = r(\eta_{\alpha^R}(T))$. It remains to show that there are s_L and s_R with $s = s_L + s_R$ such that $(T_L, s_L) \in \mathcal{R}(b_L)$ and $(T_R, s_R) \in \mathcal{R}(b_R)$.

Let $T'_L = r(\eta_{\alpha^L}(\eta_{\alpha^{\rightarrow R}}(\eta_{\alpha^s_{b_R}}(T'))))$ and $T'_R = r(\eta_{\alpha^R}(\eta_{\alpha^{\rightarrow L}}(\eta_{\alpha^s_{b_L}}(T'))))$.

Note that $T_L = r(\eta_{\alpha^s_{b_L}}(T'_L))$ because of Lemma **??** and the observation that $\eta_{\alpha^s_{b_L}} \circ \eta_{\alpha^L} \circ \eta_{\alpha^{\rightarrow R}} \circ \eta_{\alpha^s_{b_R}} = ?$.

Towards showing the reverse direction, suppose that our construction adds the record $(\hat{T}, s_L + s_R)$ and let T, T_L, T_R be as defined in the construction. Recall that:

- $\hat{T}$ is reduced and $\hat{T} = \eta_{\alpha^{LR\rightarrow}}(T)$,
- $T_L = r(\eta_{\alpha^L}(T))$ and $(T_L, s_L) \in \mathcal{R}(b_L)$,
- $T_R = r(\eta_{\alpha^R}(T))$ and $(T_R, s_R) \in \mathcal{R}(b_R)$.

Let T'_L be the reduced DT template for b_L such that $T_L = r(\eta_{\alpha^s_{b_L}}(T'_L))$ and $s_L = |V(T'_L) \setminus V(T_L)|$, which exists because $(T_L, s_L) \in \mathcal{R}(b_L)$. Similarly, let T'_R be the reduced DT template for b_R such that $T_R = r(\eta_{\alpha^s_{b_L}}(T'_R))$ and $s_R = |V(T'_R) \setminus V(T_R)|$, which exists because $(T_R, s_R) \in \mathcal{R}(b_R)$.

We now show how to construct a witness T' (from T, T'_L, and T'_R) for the validity of the record $(\hat{T}, s_L + s_R)$, i.e., T' is a reduced DT template for b such that $\hat{T} = r(\alpha^s_b(T'))$ and $s_L + s_R = |V(T') \setminus V(\hat{T})|$.

todo:show minimality here? maybe it can be done using the forward direction?

Suppose that there is a reduced DT template T' for b such that $\hat{T} = r(\alpha^s_b(T'))$ and $|V(T') \setminus V(\hat{T})| < s_L + s_R$.

Informally, we obtain T' from T after reversing the relabelling and reduction operations applied to T'_L and T'_R to obtain T_L and T_R, respectively; recall that $T_H = r(\eta_{\alpha^s_{b_H}}(T'_H))$ for

$H \in \{L, R\}$. That is, we will reverse the labelling for the nodes in T and add back the nodes to T that have been removed from T'_L and T'_R.

Let $H \in \{L, R\}$. Because T_H is obtained from T by reduction, every node in T_H corresponds to a unique node in T. Therefore, there is an injective function $x_H : V(T_H) \to V(T)$ mapping every node in T_H to its original node in T. Similarily, because T_H is obtained from T'_H by reduction, there is an injective function $y_H : V(T_H) \to V(T'_H)$ mapping every node in T_H to its original node in T'_H. See also Figure 2 for an illustration of these mappings.

$$\begin{tikzcd}
T'_L \arrow[r, dash] & T' \arrow[r, dash] \arrow[d, dash, thick] & T'_R \\
T_L \arrow[u, hook, "y_L"] \arrow[r, hook, "x_L"] & T & T_R \arrow[u, hook', "y_R"] \arrow[l, hook', "x_R"']
\end{tikzcd}$$

Figure 2

Our first order of business is to rename all forgotten features in T to their real features as given by T'_L and T'_R. That is, for every node t in T assigned to a forgotten feature, i.e., $\mathit{feat}(t) \in F_{[k_L]} \cup F_{[k_R]}$, we do the following. If $\mathit{feat}(t) \in F_{[k_H]}$ for $H \in \{L, R\}$, then t is also in T_H and hence also in T'_H. Therefore, we can change $\mathit{feat}(t)$ to the real feature assigned to t in T'_H. Let T^0 be the DT obtained from T after renaming all forgotten features to real features in this manner.

todo: explain

Consider an edge $e = (p, c)$ in T_L such that p is the parent of c in T_L. Then, e corresponds to a path $P'_L(e)$ between $y_L(p)$ and $y_L(c)$ in T'_L. Similarily, e corresponds to a path $P_L(e)$ between $x_L(p)$ and $x_L(c)$ in T^0.

Our next order of business is now to add all nodes to T^0 that have been removed when going from T'_L to T_L (via the reduction $r(\eta_{\alpha^s_{b_L}}(T'_L))$). To achieve this, we go over every edge $e = (p, c)$ of T_L such that p is the parent of c in T_L and plugin the path $P'_L(e)$ (from T'_L) into the last edge on the path $P_L(e)$ (from T^0). Let T^1 be the tree obtained from T^0 after doing this operation for every edge of T_L.

Consider an edge $e = (p, c)$ in T_R such that p is the parent of c in T_R. Then, e corresponds to a path $P'_R(e)$ between $y_R(p)$ and $y_R(c)$ in T'_R. Similarily, e corresponds to a path $P_R(e)$ between $x_R(p)$ and $x_R(c)$ in T^1. Similarily to above, we now add all nodes to T^1 that have been removed when going from T'_R to T_R (via the reduction $r(\eta_{\alpha^s_{b_R}}(T'_R))$). To achieve this, we go over every edge $e = (p, c)$ of T_R such that p is the parent of c in T_R and plugin the path $P'_R(e)$ (from T'_R) into the last edge on the path $P_R(e)$ (from T^1). Let T' be the tree obtained from T^1 after doing this operation for every edge of T_R.

We now show that T' is indeed a witness for the validity of the record $(\hat{T}, s_L + s_R)$, i.e., T' is a reduced DT template for b such that $\hat{T} = r(\alpha^s_b(T'))$ and $s_L + s_R = |V(T') \setminus V(\hat{T})|$.

We start by showing that $\hat{T} = r(\eta_{\alpha^s_b}(T'))$. Because $\hat{T} = \alpha^s_b(T^0)$, it suffices to show that the only nodes removed from T' are the ones that we added to T^0 to obtain T'. Or in other words, we need to show that only the nodes that are redundant in $\eta_{\alpha^s_b}(T')$ are the nodes in $V(T') \setminus V(T^0)$.

Consider a node $t \in V(T') \setminus V(T^0)$, i.e., t is a node that we added to T^0 to obtain T'. Then, $t \in V(T'_H) \setminus V(T_H)$ for some $H \in \{L, B\}$. Because $T_H = r(\eta_{\alpha^s_{b_H}}(T'_H))$, t is redundant in $\eta_{\alpha^s_{b_H}}(T'_H)$, because of some node $t' \in V(T_H)$ with $\alpha^s_{b_H}(\mathit{feat}_{T'_H}(t)) = \alpha^s_{b_H}(\mathit{feat}_{T'_H}(t'))$. Since $t' \in V(T_H)$ also $t' \in V(T')$ and therefore t is also redundant in $\eta_{\alpha^s_b}(T')$ (because of t'), as required.

Now consider a node $t \in V(T^0)$ and assume for a contradiction that t is redundant in $\alpha_b^s(T')$ because of some node $t' \in V(T')$ with $\alpha_b^s(\mathit{feat}_{T'}(t)) = \alpha_b^s(\mathit{feat}_{T'}(t'))$. Then, because $\hat{T} = \alpha_b^s(T^0)$ is reduced, we obtain that $t' \in V(T') \setminus V(T^0)$. Therefore, $t' \in V(T'_H) \setminus V(T_H)$ for some $H \in \{L, R\}$. But then, t' is redundant in $\eta_{\alpha_b^s}(T'_H)$ because of some node $t'' \in V(T_H)$ with $\alpha_b^s(\mathit{feat}_{T'}(t'')) = \alpha_b^s(\mathit{feat}_{T'_H}(t'))$, which implies that also t is redundant in $\hat{T}$ because of t'' a contradiction to our assumption that $\hat{T}$ is reduced. This shows that $\hat{T} = r(\eta_{\alpha_b^s}(T'))$. Moreover, because $|V(T^0)| = |V(\hat{T})|$ and $|V(T') \setminus V(T^0)| = s_L + s_R$, it also follows that $s_L + s_R = |V(T') \setminus V(\hat{T})|$.

proof of minimality of T' is still missing

Moreover, $V(T) \setminus Im(x_H)$ and $V(T'_H) \setminus Im(y_H)$ can be partitioned into subtrees that have been deleted after the application of $r \circ p_*$, $r \circ p'_*$ on T or of the standard reduction on T'_H: let X_H^* and Y_H^* be the set of roots of the above subtrees in $V(T) \setminus Im(x_H)$ and $V(T'_H) \setminus Im(y_H)$ respectively. In addition, for every element $y \in Y_H^*$, let Y_y^H be the maximal subtree of T'_H rooted at y with no elements from $Im(y_H)$ and that does not contain any vertex from $Y_H^* \setminus \{y\}$; let (Y_y^H, S_y^H) the corresponding single pair. In a similar way, for every element $x \in X_H^*$, let X_x^H be the maximal subtree of T rooted at x with no elements from $Im(x_H)$ and that does not contain any vertex from $X_H^* \setminus \{x\}$; let (X_x^H, S_x^H) the corresponding single pair. Finally, for every $y \in Y_H^*$, let P_y^H be the shortest downwards path in T'_H that contains y and with both endpoints in $Im(y_H)$, say $y_H(t)$ and $y_H(t')$.

Claim 1: For every $H \in \{L, R\}$ and for every $y, y' \in Y_H^*$, the paths P_y^H and $P_{y'}^H$ are either edge disjoint or $P_y^H = P_{y'}^H$.

Proof. If P_y^H and $P_{y'}^H$ are edge disjoint, then the statement is proven immediately. Suppose P_y^H and $P_{y'}^H$ share an edge. By minimality and the fact they are downwards paths, P_y^H and $P_{y'}^H$ share the endpoint towards the root. If they also share the other endpoint, then the statement is proven immediately. Suppose now their endpoints towards the leaves is different, say w and w', and consider the last edge those paths have in common in a root-to-leaf order, say uv.

Without loss of generality, we can assume w belongs to the left branch of v and w' belongs to the right branch of v. Note that $v \in V(T'_H) \setminus Im(y_H)$, or we get a contradiction due the minimality of P_y^H. Now we get the following contradiction: by construction, w and w' are both elements of $Im(y_H)$ but at least one of them must be in $V(T'_H) \setminus Im(y_H)$ since it is an element of either Y_y^H or of $Y_{y'}^H$. This proves Claim 1.

Now for every $y \in Y_H^*$ we consider the path Q_y^H in T having endpoints $x_H(t)$ and $x_H(t)$.

Now we are able to describe how to obtain a witness T' of T for b. For every $y \in Y_L^*$, in the last edge of path Q_y^L we plug in the single pair $(Y_{y'}^L, S_{y'}^L)$ rooted at y', for every internal node y' of P_y^L, in the order the nodes y' apprear in P_y^L. Note that, in the case an element of Y_L^* is present in more than one P_y^L, we plug in the corresponding single pair only once. Note also that whenever we plug in some single pair (Y_y^L, S_y^L) in a DT, the tree Y_y^L has real features and future features as nodes. Call this graph T^*. Now we do the same sequence of plug ins of the single pairs corresponding to the internal vertices of P_y^R in the last edge of the path Q_y^R. Again, in the case an element of Y_R^* is present in more than one P_y^R, we plug in the corresponding single pair only once. Call the tree obtained in this way T'. Node that T' contains real features from $\mathit{feat}(b_L)$ and from $\mathit{feat}(b_R)$ and future features with labels in $\mathcal{P}([k])$.

To conclude this part of the proof we have to show two things: (i) T is obtained from T' after removing s vertices; (ii) T' is a real DT for b. We start proving (i): by construction T' is obtained from T after adding s_L elements from T'_L and s_R elements from T'_R, and so with $s_L + s_R = s$ more elements.

Before considering statement (ii), we consider the following relabelling p_+ of T': every real feature in $feat(b_R)$ is a assigned to a feature with its label at node b_R and every other feature is assigned to itself. The real DT T'_L can be obtained from T' by the application of the composition $r \circ p_* \circ p_+$.

Now we consider statement (ii). We show that given an example $e \in exam(b_L)$, e is correctly classified by T' and to do so we show that e ends in a leaf of T' that corresponds to the leaf where e ends in T'_L. Say that e goes along a path P of T'_L from the root to a leaf ℓ and let Q be the corresponding path in T', i.e. the path from r to ℓ (note that by construction ℓ is present in T' and is still a leaf). Let v be a node of Q, we can have the following different cases.

- v is a real feature from $feat(b_L)$: v is also present in T'_L as real feature;
- v is a real feature from $feat(b_R)$: v might not be present in T'_L due reductions but if it is present it is a future feature A_i for some $i \in [k]$;
- v is a future feature f_A: v might not be present in T'_L due reductions but if it is present it is still the same future feature A_i.

If v is present in T'_L then the behaviour of v on e in T'_L and in T' is the same. Suppose now v is a node of Q that is being reduced due his label and so it is not present in T'_L. This means there is a set of ancestors of v such that their labels allows to remove v and by construction v behaves on e like those ancestors. This proves e goes along Q and in particular it ends at leaf ℓ and so T' is a real DT for b_L. With symmetric construction, we show that T' is also a real DT for b_R.

Now we prove the backward direction. Let T be a reduced DT such that s is the minimum number of elements that have been deleted from a witness T' of T for b. In particular, we recall that T' is a real DT for b with actual feature labels in $[k] \cup [k']$ and future feature labels in $\mathcal{P}([k])$.

We create at real DT T'_L by the application of the composition $r \circ p_* \circ p_+$ to T'. By assumption T' is a real DT for b_L and by construction T'_L is a real DT for b_L. Denote with T_L the DT template obtained from T'_L by standard reduction and denote with s_L the number of nodes that have been deleted from T'_L to obtain T. By induction we have $(T_L, s_L) \in \mathcal{R}(b_L)$. Now we note that T_L is obtained from T after the application of the composition $r \circ p_*$. In a symmetric way, we construct T'_R, T_R and the record $(T_R, s_R) \in \mathcal{R}(b_R)$. Then $(T, s_L + s_R) \in \mathcal{R}(b)$. ◀

▶ **Lemma 14** (relabel node). *Let $b \in V(B)$ be relabel node. Then $\mathcal{R}(b)$ can be computed in time $\mathcal{O}(k(2k + 2^k + 2)2^{3k+1})$.*

Proof. Let b_C be the unique child of b in B. Let R be the mapping of $[k]$ to itself that represent the node b. Moreover, since we are considering a *nice* NLC-expression we can assume R is the identity mapping, i.e. $R(\ell) = \ell$, for all values except for a unique element i of its domain, i.e. $R(i) = j$ for some $j \in [k] \setminus \{i\}$.

We say that a future feature A is *good* if it does not distinguish between i and j, that is $i \in A$ if and only if $j \in A$, and *bad* otherwise. Let (T_C, s_C) be an element of $\mathcal{R}(b_C)$. Let p'' the following relabelling of the DT template T_C: every feature with label i is assigned to label j and every future feature with label A is assigned to the future feature with label $A \setminus \{i\}$.

If T_C has a bad future feature then we do not take any other action. Suppose now T_C has only good future features; now let T be the DT template obtained from T_C after the application of the composition $r \circ p''$ and let s^* be the number of nodes that have been deleted from T_C to T.

If there is a record in $\mathcal{R}(b)$ of the form (T, s') for some integer $s' \leq s_C + s^*$ then we do not take any other action. If there is a record in $\mathcal{R}(b)$ of the form (T, s') for some integer $s' > s_C + s^*$ then we replace it with $(T, s_C + s^*)$. If there is no record in $\mathcal{R}(b)$ of the form (T, s') for some integer s' then we add $(T, s_C + s^*)$ to $\mathcal{R}(b)$.

Now we want to evaluate the running time of computing $\mathcal{R}(b)$. Consider record (T_C, s_C) in $\mathcal{R}(b_C)$. In $\mathcal{O}(k)$ time we check if T_C all the future features are good. For every such DT T_C, there are at most 2^{2k} paths from the root to the leaves and for every of these paths there are at most k nodes for each of the following: feature with label i and and future feature that contains i. This means $r \circ p''$ can be done in $\mathcal{O}(k)$ time. This means to compute $\mathcal{R}(b)$ takes $\mathcal{O}(k|\mathcal{R}(b_C)|) = \mathcal{O}(k(2k + 2^k + 2)2^{3k+1})$ time.

Now we have to show the correctness of the construction for $\mathcal{R}(b)$, i.e. $(T, s) \in \mathcal{R}(b)$ if and only if s is the minimum number of elements that have been deleted from a witness T' of T for b.

We start with the forward direction. Let $(T, s) \in \mathcal{R}(b)$. By construction there exists a record $(T_C, s_C) \in \mathcal{R}(b_C)$ such that T is obtained from T_C after the application of $r \circ p''$ and let $s^* = s - s_C$. By induction s_C is the minimum amount of nodes that have been deleted from a witness T'_C of T_C for b_C. By construction we also know that every future feature of both T'_C and T_C is good.

Denote with T' the real DT obtained T'_C after the application of $r \circ p''$: note that this last reduction does not any node since every future feature of T'_C is good and there is no feature with label i. To conclude this part of the proof we have to show two things: (i) T is obtained from T' after removing s vertices; (ii) T' is a witness of T for b.

Before proving (i), we describe how T can be obtained from T'. Let p''' be the following relabelling of T': every real feature that contains j is assigned to the real feature $A \cup \{i\}$ and every other feature is assigned to itself. Then the application of the composition p''', the standard reduction and $r \circ p''$ to T' is exactly the standard reduction for T' which then result to the DT template T. By Lemma 15 the score of the standard reduction from T' to T is exactly $s_C + s^* = s$.

Now we consider statement (ii). First note that $exam(b) = exam(b_C)$. We show that a given example $e \in exam(b)$ is correctly classified by T'. Say that e goes along a path P of T'_C from the root to a leaf ℓ. We show e goes along the path P in T' as well: every real feature has not changed and so e behaves the same. Since every future feature of T'_C is good, then e behave the same on the corresponding future feature of T'.

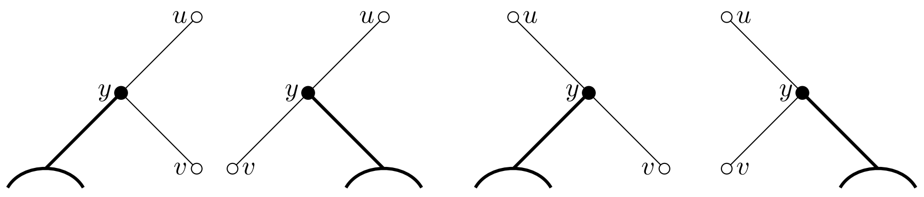
Now we prove the backward direction. Let T be a reduced DT such that s is the minimum number of elements that have been deleted from a witness T' of B for b. In particular, we recall that real T' is a DT for b with real features and future feature labels in $\mathcal{P}([k] \setminus \{i\})$.

We create the real DT T'_C as the application of $r \circ p'''$ to T', the DT template T_C as the application of the standard reduction to T'_C. By construction we have $(T_C, s_C) \in \mathcal{R}(b_C)$, where s_C is the number of nodes that have been removed from T'_C to T_C. Note that T_C has only good future features. Finally we note that T is obtained from T_C by the application of $r \circ p''$. ◀

3.4 Formal Definition of Records and Preliminary Results

We start off with some definitions. We say an edge is a *left (right) edge* of a subcubic rooted tree if it connects a non-leaf node with his left (resp. right) child. Let Y be a rooted subcubic tree and $S \in \{left, right\}$, then we say the pair (Y, S) is a *single pair* if the root of Y has at most one child and the side S indicates whether the edge from the root is either a left or right edge. Moreover, we say that (Y, S) is single pair in a subcubic rooted tree T if Y is a maximal subtree of T and in Y the root have at most the S child. Note that when tree of a single pair is made of just a node, the side is not relevant.

Now we can define two operations on subcubic rooted trees and single pairs. We say that we *plug in* a single pair (Y, S) in a left (right) edge uv as follows: we make the root y of Y the left (right) child of u, $Y \setminus \{y\}$ to be the S subtree of y and v to be the $H \in \{left, right\} \setminus S$ child of y. See Figure 3 for the corresponding drawings. Note after a plug in of a single pair in an edge, the node v belongs in the same side of the subtree rooted at u as it was before the plug in.



(a) $(Y, left)$ left. **(b)** $(Y, right)$ left. **(c)** $(Y, left)$ right. **(d)** $(Y, right)$ right.

Figure 3 The drawings describe the plug in operation in the different four cases. The bold part highlight the single pair (Y, S).

Let (Y, S) be a single pair in a rooted subcubic tree T, then we *remove* (Y, S) from T as follows. Let y be the root of Y. If y is the root of T, then we obtain an empty tree. If y is a leaf node of T, then we obtain $T - y$. Otherwise let y be a non-root and non-leaf node, let u be the parent of y and v be the child of y that is not in $V(Y)$, then we consider the tree obtained from T after replacing y with v as the child of u and deleting Y. See Figure 4 for an example.

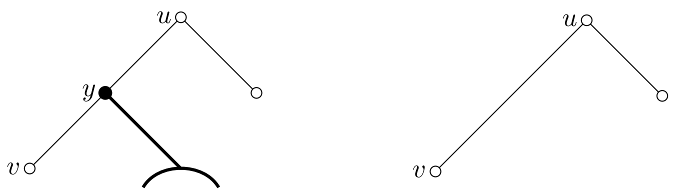


Figure 4 The drawing describe an example of the remove operation: a single pair $(Y, right)$ is removed from a subcubic rooted tree. The bold part highlight the single pair (Y, S).

It is clear from the four different plug in cases that if we want to plug in two pairs (Y, S) and (Y', S') on an edge uv such that the ancestor-descendant relationship is given, say y of Y has to be in the path from the root to y' of Y', then we can do these plug ins in any order but with some care. It is the same if we first plug in (Y, S) in the edge uv and then plug in (Y', S') in the edge yv or if we first plug in (Y', S') in the edge uv and then plug in (Y, S) in the edge uy'. See Figure 5 for the an example.

For a subset of labels $A \subseteq [k]$, we define the feature template f_A by setting $e(f_A) = 1$ if and only if $lab(e) \in A$ and $e(f_A) = 0$ otherwise. With a small abuse of notation, we often identify the feature template f_A with the corresponding subset of labels A.

Suppose we have a DT such that some feature label i occurs twice on a path from the root to the leaves, say f_1 is the instance closer to the root and f_2 is the other instance. If f_2

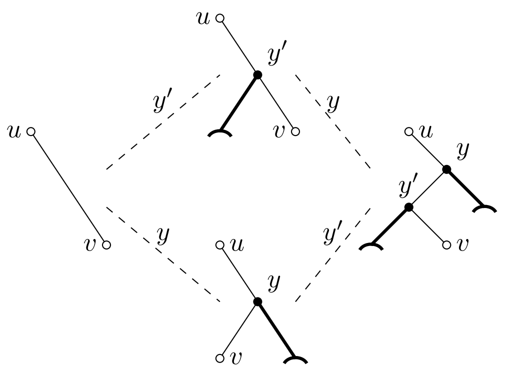


Figure 5 An example of plugging in two pairs $(Y, left)$ and $(Y', right)$ in a left edge uv.

is in the left (resp. right) subtree of f_1, we remove f_2's right (resp. left) subtree. In this case we say we have done an *actual removal*.

Suppose we have a feature template labelled A in our decision tree. Let $A_1, \ldots, A_\ell$ be the sequence of feature templates on the path from the root to A in order (not including A). Let $A'_i = A_i$ if A is in the right sub-tree of A_i and let $A'_i = \overline{A_i}$ otherwise. If $\overline{A} \subseteq A'_1 \cup \ldots \cup A'_\ell$, then we remove the subtree rooted at the left child of A. If $A \subseteq \overline{A'_1} \cup \ldots \cup \overline{A'_\ell}$, then we remove the subtree rooted at the right child of A. In this case we say we have done a *template removal*. If this procedure has been applied to a record exhaustively, we say that the DT is *reduced*.

To be short, for a DT T and a node v, we write $v \in T$ instead of $v \in V(T)$ and $v \notin T$ otherwise. In a DT T we say that path p is a *downward* pard path if it is contained in a path having the root as endpoint.

We now formally define two important operations. Given a DT T, we say that we *reduce* T if we exhaustively do actual removals and template removals. Call $r(T)$ the resulting DT.

Recall that in any DT T, every non-leaf node v has one of the following three contents: v is a real feature (without label), or v is a feature with a label, or v is a future feature with the corresponding subset of labels. A *relabelling* p for T is an assignment of contents of T as follows. Every feature is assigned to a feature with is either future, real or with a label. We say that we *relabel* the DT T via the relabelling p if for every node of T we apply the corresponding assignment and call $p(T)$ the resulting DT.

The following lemma shows that, after repeatedly applying it the necessary amount of times, to obtain a reduced DT after a sequence of relabels, it is safe to reduce at the end.

▶ **Lemma 15** (Relabelling Lemma). *Let T be a DT and p be relabelling of T. Then $(r \circ p \circ r)(T) = (r \circ p)(T)$.*

Proof. For every $v \in T$, we want to prove $v \in (r \circ p \circ r)(T) \Leftrightarrow v \in (r \circ p)(T)$.

$\Rightarrow$ Suppose there is a node $v \notin (r \circ p)(T)$. Since $v \in p(T)$, there is a set of ancestors of v in $p(T)$ that allows to remove v. Let A_v be the union of all the minimal set of ancestors of v in $p(T)$ that allows to remove v. If A_v is a set of ancestors of v in T that allows to reduce v then $v \notin r(T)$ and so $v \notin (r \circ p \circ r)(T)$. Otherwise let A'_v be the subset of A_v in $(p \circ r)(T)$. We conclude by noting that A'_v contains one of the minimal sets A_v is composed of and so $v \notin (r \circ p \circ r)(T)$.

$\Leftarrow$ Suppose there is a node $v \notin (r \circ p \circ r)(T)$. If $v \in (p \circ r)(T)$, there exists a set A_v of ancestors of v in $(p \circ r)(T)$ that allows to reduce v. Then A_v is a set of ancestors of v in $p(T)$ that allows to reduce v and so $v \notin (r \circ p)(T)$. If $v \notin (p \circ r)(T)$ then $v \notin r(T)$: there exists a set A_v of ancestors of v in T that allows to remove v. This means A_v is a set of ancestors of v in $p(T)$ that allows to remove v and so $v \notin (r \circ p)(T)$. ◀

We say that a DT T is a *real DT* if every non-leaf node is either a real feature or a future feature, whereas it is a *DT template* if it contains no real feature.

Let B be a rooted subcubic tree that corresponds to a k-NLC expression of the graph $G_I(E)$. For $b \in V(B)$, we write $feat(b)$ and $exam(b)$ for the sets of features and examples introduced at nobe b. We say that a real DT T is a DT for the node b if every real feature of T is an element of $feat(b)$ and every example in $exam(b)$ is correctly classified by T, i.e. if $e \in exam(b) \cap E^+$ then e ends in a leaf with a $+$ label and if $e \in exam(b) \cap E^-$ then e ends in a leaf with a $-$ label.

Given a real DT T and a node $b \in B$, often we want to perform a very specific composition of operations. Let p_b be the following relabelling of T: every real feature of T is assigned to a feature with the label given by the k-NLC expression at node b and every other feature is assigned to itself. Then the composition $r \circ p_b$ is called the *standard reduction* of T at node b. Given a DT T and a node $b \in B$, it is useful to give the following relabelling p'_b: every feature with a label is assigned to the real feature of that node. The relabelling p'_b is called the *real relabelling* of T at node b.

We say that a DT template T is a DT for the node b if there exits a real DT T' for b such that T is the standard reduction of T'. In this case we say that T' is the witness of T for b.

▶ **Lemma 16.** *If there are ℓ features with labels and 2^h future features, then every reduced DT template has height at most $\ell + h$. Furthermore, every path from the root to the leaves contains at most ℓ features with label and at most $h - 1$ future features.*

Proof. Consider a path P of maximum length from the root to the leaves in a reduced DT template T. By the assumptions on T, no feature with label appears more than once on this path: the number of these feature nodes on this path is at most ℓ. Consider two future features f_A and $f_{A'}$ that appear in P, say f_A is the instance closer to the root. Since T is reduced, we must have that $\emptyset \subset A' \subset A$. Since the label of any future feature has at most h elements, there can be at most $h - 1$ feature template nodes on this path. The path ends with a leaf node, so this gives a total of $\ell + h - 1 + 1 = \ell + h$ nodes, as required. ◀

▶ **Lemma 17.** *If there are ℓ features with label and 2^h future features, then there are at most $(\ell + 2^k + 2)2^{\ell+k+1}$ reduced DT templates. Furthermore, these can be enumerated in $\mathcal{O}((\ell + 2^k + 2)2^{\ell+k+1})$-time.*

Proof. By Lemma 16, the tree has height at most $\ell + k$. Each node of the decision tree could be a feature with label, a future feature, or a leaf: at most $\ell + 2^h + 2$ different contents. Since there are at most $2^{\ell+h+1}$ nodes in the tree, there are at most $(\ell + 2^h + 2)2^{\ell+h+1}$ possible decision trees. ◀

The *semantics* for a record are defined as follows. We say that a pair (T, s) is a *record* for the node $b \in B$ and we write $(T, s) \in \mathcal{R}(b)$, if T is a DT template for b and s is the minimum number of elements that have been deleted from a witness T' of T for b.

3.5 Proof to the Main Result

Now, it suffices to compute $\mathcal{R}(b)$ via leaf-to-root dynamic programming. The following four lemmas show how this can be achieved for all of the four types of nodes in a k-NLC expression tree B.

▶ **Lemma 18** (leaf node). *Let $b \in V(B)$ be a leaf node. Then $\mathcal{R}(b)$ can be computed in time $\mathcal{O}(k(2^k+3)2^{k+2})$.*

Proof. Let v be the vertex of $G_I(E)$ that corresponds to the leaf node b. This means either $v \in E$ or $v \in feat(E)$.

We have to enumerate all possible reduced DT templates T for b. It is enough to consider all reduced DT templates T of height at most $k+1$ and discard those that are not DT templates for b; these can be enumerated in time $\mathcal{O}((2^k+3)2^{k+2})$ by Lemma 17 and the check can be done in time $\mathcal{O}(k)$. We add the pair $(T, 0)$ to the set of records $\mathcal{R}(b)$.

Now we have to show the correctness of the construction for $\mathcal{R}(b)$, i.e. $(T, s) \in \mathcal{R}(b)$ if and only if s is the minimum number of elements that have been deleted from a witness T' of T for b.

We start with the forward direction. Let $(T, s) \in \mathcal{R}(b)$. By construction, we have that $s = 0$ and T is a DT template for b which is already reduced. Then T is trivially a witness of T for b.

Now we prove the backward direction. Let T be a reduced DT template such that 0 is the minimum number of elements that have been deleted from a witness T' of T for b. This means T' is obtained from T after the real relabelling at node b is applied: T is a DT template among the considered DTs above which leads to the fact that $(T, 0) \in \mathcal{R}(b)$. ◀

▶ **Lemma 19** (join node). *Let $b \in V(B)$ be a join node. Then $\mathcal{R}(b)$ can be computed in time $\mathcal{O}(k(2k+2^k+2)2^{6k+1})$.*

Proof. Let b_L and b_R be the left, resp. right, child of b in B: we may assume the labels for $feat(b_L)$ are in $[k]$ and the labels for $feat(b_R)$ are in $[k']$. Moreover, let M be the $k \times k$ $\{0, 1\}$ matrix that represent the node b. Finally, for every label $i \in [k]$, let $A_i = \{j \in [k] \mid M_{i,j} = 1\}$.

We consider every reduced DT T for b with feature labels in $[k] \cup [k']$ and future feature labels in $\mathcal{P}([k])$; these can be enumerated in time $\mathcal{O}((2k+2^k+2)2^{3k+1})$ by Lemma 17.

For every such DT T, we create a DT T_L as follows. Let p_* be the following relabelling: for every $i' \in [k']$, every feature with label i' is assigned to the future feature A_i. Then we apply the composition $r \circ p_*$ to T. In a symmetrical way we create a DT T_R. Let p'_* be the following relabelling: for every $i \in [k]$, every feature with label i is assigned to the future feature $A_{i'}$ and every future feature A_i is assigned to the future feature $A_{i'}$. Then we apply the composition $r \circ p'_*$ to T.

Now we want to understand if there is a record in $\mathcal{R}(b_L)$ of the form (T_L, s_L) for some positive integer s_L and if there is a record in $\mathcal{R}(b_R)$ of the form (T_R, s_R) for some positive integer s_R: if the answer is yes in both cases, we add a record $(T, s_L + s_R)$ to $\mathcal{R}(b)$; otherwise we discard this option.

Now we want to evaluate the running time of computing $\mathcal{R}(b)$. Every reduced DT T can be enumerated in time $\mathcal{O}((2k+2^k+2)2^{3k+1})$ by Lemma 17. For every such DT T, there are at most 2^{3k} paths from the root to the leaves and for every of these paths there are at most k nodes for each of the following: features with label in $[k]$, features with label in $[k']$ and future features by Lemma 16. This means $r \circ p_*$ and $r \circ p'_*$ can be done in $\mathcal{O}(k2^{3k})$ time.

Now we have to show the correctness of the construction for $\mathcal{R}(b)$. We start with the forward direction. Let $(T, s) \in \mathcal{R}(b)$. By construction there exist records $(T_L, s_L) \in \mathcal{R}(b_L)$ and $(T_R, s_R) \in \mathcal{R}(b_R)$ such that T_L and T_R are obtained by the application of $r \circ p_*$ and $r \circ p'_*$ respectively to T and $s_L + s_R = s$.

By induction, for $H \in \{L, R\}$, we know that s_H is the minimum number of elements that have been deleted from a witness T'_H of T_H for b_H.

For $H \in \{L, R\}$, we define maps x_H and y_H as follows. Let $x_H : V(T_H) \to V(T)$ and $y_H : V(T_H) \to V(T'_L)$ be the functions that maps every node of T_H to the corresponding node in T and in T'_L and note that by constructions both these maps are injective.

$$\begin{tikzcd}
T'_L \arrow[r, dash] & T' \arrow[r, dash] \arrow[d, dash, very thick] & T'_R \\
T_L \arrow[u, hook, "y_L"] \arrow[r, hook, "x_L"] & T & T_R \arrow[u, hook', "y_R"] \arrow[l, hook', "x_R"']
\end{tikzcd}$$

Moreover, $V(T) \setminus Im(x_H)$ and $V(T'_H) \setminus Im(y_H)$ can be partitioned into subtrees that have been deleted after the application of $r \circ p_*$, $r \circ p'_*$ on T or of the standard reduction on T'_H: let X^*_H and Y^*_H be the set of roots of the above subtrees in $V(T) \setminus Im(x_H)$ and $V(T'_H) \setminus Im(y_H)$ respectively. In addition, for every element $y \in Y^*_H$, let Y^H_y be the maximal subtree of T'_H rooted at y with no elements from $Im(y_H)$ and that does not contain any vertex from $Y^*_H \setminus \{y\}$; let (Y^H_y, S^H_y) the corresponding single pair. In a similar way, for every element $x \in X^*_H$, let X^H_x be the maximal subtree of T rooted at x with no elements from $Im(x_H)$ and that does not contain any vertex from $X^*_H \setminus \{x\}$; let (X^H_x, S^H_x) the corresponding single pair. Finally, for every $y \in Y^*_H$, let P^H_y be the shortest downwards path in T'_H that contains y and with both endpoints in $Im(y_H)$, say $y_H(t)$ and $y_H(t')$.

Claim 1: For every $H \in \{L, R\}$ and for every $y, y' \in Y^*_H$, the paths P^H_y and $P^H_{y'}$ are either edge disjoint or $P^H_y = P^H_{y'}$.

Proof. If P^H_y and $P^H_{y'}$ are edge disjoint, then the statement is proven immediately. Suppose P^H_y and $P^H_{y'}$ share an edge. By minimality and the fact they are downwards paths, P^H_y and $P^H_{y'}$ share the endpoint towards the root. If they also share the other endpoint, then the statement is proven immediately. Suppose now their endpoints towards the leaves is different, say w and w', and consider the last edge those paths have in common in a root-to-leaf order, say uv.

Without loss of generality, we can assume w belongs to the left branch of v and w' belongs to the right branch of v. Note that $v \in V(T'_H) \setminus Im(y_H)$, or we get a contradiction due the minimality of P^H_y. Now we get the following contradiction: by construction, w and w' are both elements of $Im(y_H)$ but at least one of them must be in $V(T'_H) \setminus Im(y_H)$ since it is an element of either Y^H_y or of $Y^H_{y'}$. This proves Claim 1.

Now for every $y \in Y^*_H$ we consider the path Q^H_y in T having endpoints $x_H(t)$ and $x_H(t)$.

Now we are able to describe how to obtain a witness T' of T for b. For every $y \in Y^*_L$, in the last edge of path Q^L_y we plug in the single pair $(Y^L_{y'}, S^L_{y'})$ rooted at y', for every internal node y' of P^L_y, in the order the nodes y' apprear in P^L_y. Note that, in the case an element of Y^*_L is present in more than one P^L_y, we plug in the corresponding single pair only once. Note also that whenever we plug in some single pair (Y^L_y, S^L_y) in a DT, the tree Y^L_y has real features and future features as nodes. Call this graph T^*. Now we do the same sequence of plug ins of the single pairs corresponding to the internal vertices of P^R_y in the last edge of the path Q^R_y. Again, in the case an element of Y^*_R is present in more than one P^R_y, we plug in the corresponding single pair only once. Call the tree obtained in this way T'. Node that T' contains real features from $feat(b_L)$ and from $feat(b_R)$ and future features with labels in $\mathcal{P}([k])$.

To conclude this part of the proof we have to show two things: (i) T is obtained from T' after removing s vertices; (ii) T' is a real DT for b. We start proving (i): by construction T'

is obtained from T after adding s_L elements from T'_L and s_R elements from T'_R, and so with $s_L + s_R = s$ more elements.

Before considering statement (ii), we consider the following relabelling p_+ of T': every real feature in $feat(b_R)$ is a assigned to a feature with its label at node b_R and every other feature is assigned to itself. The real DT T'_L can be obtained from T' by the application of the composition $r \circ p_* \circ p_+$.

Now we consider statement (ii). We show that given an example $e \in exam(b_L)$, e is correctly classified by T' and to do so we show that e ends in a leaf of T' that corresponds to the leaf where e ends in T'_L. Say that e goes along a path P of T'_L from the root to a leaf ℓ and let Q be the corresponding path in T', i.e. the path from r to ℓ (note that by construction ℓ is present in T' and is still a leaf). Let v be a node of Q, we can have the following different cases.

- v is a real feature from $feat(b_L)$: v is also present in T'_L as real feature;
- v is a real feature from $feat(b_R)$: v might not be present in T'_L due reductions but if it is present it is a future feature A_i for some $i \in [k]$;
- v is a future feature f_A: v might not be present in T'_L due reductions but if it is present it is still the same future feature A_i.

If v is present in T'_L then the behaviour of v on e in T'_L and in T' is the same. Suppose now v is a node of Q that is being reduced due his label and so it is not present in T'_L. This means there is a set of ancestors of v such that their labels allows to remove v and by construction v behaves on e like those ancestors. This proves e goes along Q and in particular it ends at leaf ℓ and so T' is a real DT for b_L. With symmetric construction, we show that T' is also a real DT for b_R.

Now we prove the backward direction. Let T be a reduced DT such that s is the minimum number of elements that have been deleted from a witness T' of T for b. In particular, we recall that T' is a real DT for b with actual feature labels in $[k] \cup [k']$ and future feature labels in $\mathcal{P}([k])$.

We create at real DT T'_L by the application of the composition $r \circ p_* \circ p_+$ to T'. By assumption T' is a real DT for b_L and by construction T'_L is a real DT for b_L. Denote with T_L the DT template obtained from T'_L by standard reduction and denote with s_L the number of nodes that have been deleted from T'_L to obtain T. By induction we have $(T_L, s_L) \in \mathcal{R}(b_L)$. Now we note that T_L is obtained from T after the application of the composition $r \circ p_*$. In a symmetric way, we construct T'_R, T_R and the record $(T_R, s_R) \in \mathcal{R}(b_R)$. Then $(T, s_L + s_R) \in \mathcal{R}(b)$. ◀

▶ **Lemma 20** (relabel node). *Let $b \in V(B)$ be relabel node. Then $\mathcal{R}(b)$ can be computed in time $\mathcal{O}(k(2k + 2^k + 2)2^{3k+1})$.*

Proof. Let b_C be the unique child of b in B. Let R be the mapping of $[k]$ to itself that represent the node b. Moreover, since we are considering a *nice* NLC-expression we can assume R is the identity mapping, i.e. $R(\ell) = \ell$, for all values except for a unique element i of its domain, i.e. $R(i) = j$ for some $j \in [k] \setminus \{i\}$.

We say that a future feature A is *good* if it does not distinguish between i and j, that is $i \in A$ if and only if $j \in A$, and *bad* otherwise. Let (T_C, s_C) be an element of $\mathcal{R}(b_C)$. Let p'' the following relabelling of the DT template T_C: every feature with label i is assigned to label j and every future feature with label A is assigned to the future feature with label $A \setminus \{i\}$.

If T_C has a bad future feature then we do not take any other action. Suppose now T_C has only good future features; now let T be the DT template obtained from T_C after the application of the composition $r \circ p''$ and let s^* be the number of nodes that have been deleted from T_C to T.

If there is a record in $\mathcal{R}(b)$ of the form (T, s') for some integer $s' \leq s_C + s^*$ then we do not take any other action. If there is a record in $\mathcal{R}(b)$ of the form (T, s') for some integer $s' > s_C + s^*$ then we replace it with $(T, s_C + s^*)$. If there is no record in $\mathcal{R}(b)$ of the form (T, s') for some integer s' then we add $(T, s_C + s^*)$ to $\mathcal{R}(b)$.

Now we want to evaluate the running time of computing $\mathcal{R}(b)$. Consider record (T_C, s_C) in $\mathcal{R}(b_C)$. In $\mathcal{O}(k)$ time we check if T_C all the future features are good. For every such DT T_C, there are at most 2^{2k} paths from the root to the leaves and for every of these paths there are at most k nodes for each of the following: feature with label i and and future feature that contains i. This means $r \circ p''$ can be done in $\mathcal{O}(k)$ time. This means to compute $\mathcal{R}(b)$ takes $\mathcal{O}(k|\mathcal{R}(b_C)|) = \mathcal{O}(k(2k + 2^k + 2)2^{3k+1})$ time.

Now we have to show the correctness of the construction for $\mathcal{R}(b)$, i.e. $(T, s) \in \mathcal{R}(b)$ if and only if s is the minimum number of elements that have been deleted from a witness T' of T for b.

We start with the forward direction. Let $(T, s) \in \mathcal{R}(b)$. By construction there exists a record $(T_C, s_C) \in \mathcal{R}(b_C)$ such that T is obtained from T_C after the application of $r \circ p''$ and let $s^* = s - s_C$. By induction s_C is the minimum amount of nodes that have been deleted from a witness T'_C of T_C for b_C. By construction we also know that every future feature of both T'_C and T_C is good.

Denote with T' the real DT obtained T'_C after the application of $r \circ p''$: note that this last reduction does not any node since every future feature of T'_C is good and there is no feature with label i. To conclude this part of the proof we have to show two things: (i) T is obtained from T' after removing s vertices; (ii) T' is a witness of T for b.

Before proving (i), we describe how T can be obtained from T'. Let p''' be the following relabelling of T': every real feature that contains j is assigned to the real feature $A \cup \{i\}$ and every other feature is assigned to itself. Then the application of the composition p''', the standard reduction and $r \circ p''$ to T' is exactly the standard reduction for T' which then result to the DT template T. By Lemma 15 the score of the standard reduction from T' to T is exactly $s_C + s^* = s$.

Now we consider statement (ii). First note that $exam(b) = exam(b_C)$. We show that a given example $e \in exam(b)$ is correctly classified by T'. Say that e goes along a path P of T'_C from the root to a leaf ℓ. We show e goes along the path P in T' as well: every real feature has not changed and so e behaves the same. Since every future feature of T'_C is good, then e behave the same on the corresponding future feature of T'.

Now we prove the backward direction. Let T be a reduced DT such that s is the minimum number of elements that have been deleted from a witness T' of B for b. In particular, we recall that real T' is a DT for b with real features and future feature labels in $\mathcal{P}([k] \setminus \{i\})$.

We create the real DT T'_C as the application of $r \circ p'''$ to T', the DT template T_C as the application of the standard reduction to T'_C. By construction we have $(T_C, s_C) \in \mathcal{R}(b_C)$, where s_C is the number of nodes that have been removed from T'_C to T_C. Note that T_C has only good future features. Finally we note that T is obtained from T_C by the application of $r \circ p''$. ◀

Now we can finally prove Theorem 4 and Theorem **??**, which we restate here.

Theorem 4 (restated). *Let E be a CI, let (B, χ) be an NLC-expression decomposition of width k for $G_I(E)$, and let s be an integer. Then, deciding whether E has a DT of size at most s is fixed-parameter tractable parameterized by k. In particular, such computation takes $\mathcal{O}()$ time.*

Proof. We start off by computing $\mathcal{R}(b)$ for every node b of B, via leaf-to-root dynamic programming. An upper bound for the running time for this step is the number of nodes of B times the maximum running time to compute the record at each node which is given by Lemmas 18, 19 and 20.

Now we look at the root node r of B. We go through all the records of $\mathcal{R}(r)$ and select a record $(T, s) \in \mathcal{R}(r)$ such that $|T| + s$ is minimum over all DTs with no future feature. ◀

Theorem ?? (restated). DTS *is fixed-parameter tractable parameterized by NLC-width.*

4 An FPT-Algorithm for bounded solution size and δ_{max}.

Given a CI E, our aim it to construct a support set S^* for E such that there is a DT T^* for E of size at most s with $feat(T^*) = S^*$.

First of all, we note that we can enumerate all the minimal support sets in fpt-time parametrized by δ_{max} as follows.

▶ **Lemma 21** ([23]). *Let E be a CI and let k be an integer. Then there is an algorithm that in time $\mathcal{O}(\delta_{max}(E)^k|E|)$ enumerates all (of the at most $\mathcal{O}(\delta_{max}(E)^k|E|)$) minimal support sets of size at most k for E.*

Let S be a support set for E. Consider a DT T, we say that a non-leaf node t of T has an *unknown feature* for S if $feat(t) \notin S$; if $feat(t) \in S$ then we say that t has a *supported feature*. If a DT T contains an unknown feature for a support set S, then it is called a *partial DT* for S. Otherwise, that is if T does not contain any unknown features for S, T is called a *supported DT* for S.

The following Lemma express that we can already solve the problem in the case we have a supported DT T for S: this is given by the fact a feature assignment α for T is known.

▶ **Theorem 22** ([23]). *Let E be a CI, $S \subseteq feat(E)$ be a support set for E, and let s and d be an integer. Then, there is an algorithm that runs in time $2^{\mathcal{O}(s^2)}\|E\|^{1+o(1)}\log\|E\|$ and computes a DT of minimum size among all DTs T with $feat(T) = S$ and $size(T) \leq s$ if such a DT exists; otherwise* `nil` *is returned.*

▶ **Lemma 23** ([23]). *Let E be a CI, let T be DT for S of depth d and let α be a feature assignment for T. Then, there is an algorithm that runs in time $\mathcal{O}(2^{d^2/2}\|E\|^{1+o(1)}\log\|E\|)$ and decides whether the pair (T, α) can be extended to a DT for E.*

4.1 Preprocess

Let T be a pseudo DT and S be a support set for E. An *S-feature assigment* of T is a mapping from a subset of the non-leaf nodes of T to S. For every $v \in V(T)$, we define the set of *expected examples* E_v as follows:

- if v is the root, then $E_v = E$;
- if v is the left child of a supported feature node v_p with feature $f_p \in S$, then $E_v = E_{v_p}[f_p \leq th_L(v_p) + 1]$;

- if v is the right child of a supported feature node v_p with feature $f_p \in S$, then $E_v = E_{v_p}[f_p > th_R(v_p) - 1]$;
- if v is a child of an unknown feature node v_p, then $E_v = E_{v_p}$.

Node that the definition of E_v is strictly related with the following: if v has a supported feature f, let c_ℓ and c_r be the left, risp. right, child of v, we define two values $th_L(t)$ and $th_R(T)$ as follows:

- let $th_L(v)$ be the maximum value in $D_E(f)$ such that every example in $E_v[f \leq th_L(v)]$ is correctly classified in T_{c_ℓ};
- let $th_R(v)$ be the minimum value in $D_E(f)$ such that every example in $E_v[f > th_R(v)]$ is correctly classified in T_{c_r}.

Before formally proving in Lemma 24 that we are able to compute E_v and $th_L(v)$, $th_R(v)$ (when v has a supported feature) for every $v \in V(T)$, we want to describe the role of E_v in the proof of Lemma 25.

Let us consider the following situation. Suppose we are trying to find a DT of minimum size for a CI E using at least the features in a given support set S. The first step would be to compute a minimum size DT T^* for E such that $feat(T^*) = S$ (note this can be done by Theorem 23). Next we analyse the case an optimal DT for E uses not only every feature from S but some additional feature: for this reason we consider partial DTs T of size at most s such that the set of supported features of T is exactly S.

To understand the properties of an unknown feature of T, we suppose $u \in V(T)$ has an unknown feature that does not distinguish any pair of examples. Clearly T can be extended to a DT for E only if the smaller T', obtained from T by removing u and the subtree of T_u that receives no example, can. So let us suppose T can not be extended to a DT for E.

Claim 1. *For every node $v \in V(T)$, the set E_v is a set of examples such that T_v can not be extended to a DT for E.*

Let us first consider the case in which v is the root of T. Here we have that $T_v = T$ and $E_v = E$ and so the claim directly follows from the assumption that T can not be extended to a DT for E. Now let v be the child of an unknown feature node v_p. Recall that, by assumption, v_p does not distinguish any pair of examples and so it is safe to assume that E_{v_p} is a set of examples that can not be classified by extending T_v.

Finally suppose now that v is the left child of a supported feature node v_p. By the maximality of $th_L(v_p)$, the set $E_v = E_{v_p}[f_p \leq th_L(v_p) + 1]$ necessarily contains examples that can not be classified by extending T_v. In a similar way, the minimality of $th_R(v)$ shows that the claim holds when v is the right child of a supported feature node. This proves the Claim 1.

Thanks to Claim 1, it follows that for every node v that has a supported feature, $th_L(v) < th_R(v)$. Suppose that for some node v with supported feature f we have $th_L(v) \geq th_R(v)$. Setting $(feat(v) = th_L(v))$ generates a DT for E_v, which is a contraddiction.

▶ **Lemma 24.** *Let E be a CI, let T be a partial DT of depth at most d and let $\overline{\alpha}$ be a S-feature assignment of T for a support set S. Then there is an algorithm that runs in time $\mathcal{O}(2^{d^2/2} n^{1+o(1)} \log n)$ and computes the set E_v and thresholds $th_L(v)$ and $th_R(v)$ for every node $v \in V(T)$.*

Proof. The idea is to use the recursive algorithm **findLR** illustrated in Algorithm 1. That is, given E, T, $\overline{\alpha}$, the algorithm **findLR** attempts to find the triples $(E_v, th_L(v), th_R(v))$ for every node $v \in V(T)$. Lines 3 to 4: if T consists of a leaf node, the algorithm just report $(E, \texttt{nil}, \texttt{nil})$. Let c_ℓ and c_r be the left, risp. right, child of the root v. Lines 6 to 11: if the root of T has an unknown feature, the algorithm calls itself recursively to compute the triple for (E, T_{c_ℓ}, α) and (E, T_{c_r}, α). Lines 13 to 15: if the root of T has a supported feature f, the algorithm computes the pair (t_ℓ, t_r) for the root using the algorithm **binarySearch** and then calls itself recutsively to compute the triple for $(E[f \leq t_\ell + 1], T_{c_\ell}, \alpha)$ and $(E[f > t_r - 1], T_{c_r}, \alpha)$.

A key element for the correctness of **findLR** is the algorithm **binarySearch** illustrated in Algorithm 2. Given E, T, α, f, c_ℓ and c_r, this algorithm computes the pair (t_ℓ, t_r) for the root of T. This sub-routine performs a standard binary search procedure on the array D containing all the values in $D_E(f)$ in ascending order to find maximum t_ℓ and minimum t_r such that (T_{c_ℓ}, α) and (T_{c_r}, α) can be extended to DT for $E[f \leq t_\ell]$ and for $E[f > t_r]$ respectively. To achieve this, the sub-routine makes at most $log|E|$ calls to **findTH**; note that each of those calls is made for a tree of smaller depth. Lines 3 to 12: the algorithm finds the maximum t_ℓ by calling algorithm **findTH** in Line 6 repeatedly. Lines 13 to 22: the algorithm finds the minimum t_r by calling algorithm **findTH** in Line 16 repeatedly.

A sub-routine used for **binarySearch** is the algorithm **findTH** illustrated in Algorithm 3. This algorithm is very similar to Algorithm 1 but the output is some way much simpler.

The running time of Algorithm 1 can now be obtained by multiplying the number of recursive calls to **findLR** with the time required for one recursive call. To obtain the number of recursive calls first note that if **findLR** is called with a pseudo DT of depth d, then it makes at most $(2\log n) + 2$ recursive calls to **findLR** with a pseudo DT of depth at most $d - 1$, where $n = |E|$. Therefore the number $T(n, d)$ of recursive calls for a pseudo DT of depth d is given by the recursion relation $T(n, d) = (2(\log n) + 2)T(n, d - 1)$ starting with $T(n, 0) = 0$. This implies that $T(n, d) \in \mathcal{O}((\log n)^d)$. Finally, the runtime for one recursive call is easily seen to be at most $\mathcal{O}(n \log n)$. Hence, the total runtime of the algorithm is at most $\mathcal{O}((\log n)^d n \log n)$, which because (see also [9, Exercise 3.18]):

$$(\log n)^d \leq 2^{d^2/2} 2^{\log\log d^2/2} = 2^{d^2/2} n^{o(1)}$$

is at most $\mathcal{O}(2^{d^2/2} n^{1+o(1)} \log n)$. ◀

4.2 The algorithm

Now we have computed a set E_v for every node $v \in V(T)$, whether it is a leaf or has a supported or unknown feature. A *pool set* for node $v \in V(T)$ is a set $\Pi(v) \subseteq E_v$, such that if every example of $\Pi(v)$ arrives at node v then either

- T can not be extended to a DT for E, or
- there is an node $u \in V(T_v)$ with unknown feature f_u and two elements $e, e' \in \Pi(v)$ such that f_u must distinguish e and e'.

For every node $v \in V(T)$, we define $\Pi(v)$ in a leaves-to-root fashion as follows:

(a) if v is a negative leaf then $\Pi(v) = \{e^+\}$, where e^+ is any example in $E^+ \cap E_v$;
(b) if v is a positive leaf then $\Pi(v) = \{e^-\}$, where e^- is any example in $E^- \cap E_v$;
(c) if v is a supported feature node. Let c_ℓ and c_r be the left, resp. right, child of v, then $\Pi(v) = \Pi(c_\ell) \cup \Pi(c_r)$;
(d) if t is an unknown feature node. Let c_ℓ and c_r be the left, resp. right, child of v, then $\Pi(v) = \Pi(c_\ell) \cup \Pi(c_r)$.

Algorithm 1 Algorithm to compute the triple $(E_v, th_L(v), th_R(v))$ for every node $v \in V(T)$.

```
Input: CI $E$, pseudo DT $T$, $S$-feature assignment $\alpha$ for $T$
Output: a triple $(E_v, th_L(v), th_R(v))$ for every node $v \in V(T)$.
 1: function findLR($E$, $T$, $\alpha$)
 2:     $r \leftarrow$ “root of $T$”
 3:     if $r$ is a leaf then
 4:         return $(E, \texttt{nil}, \texttt{nil})$
 5:     $c_\ell, c_r \leftarrow$ “left child and right child of $r$”
 6:     if $r$ has an unknown feature then
 7:         $\lambda_\ell \leftarrow$ FINDLR($E$, $T_{c_\ell}$, $\alpha$)
 8:         $\lambda_r \leftarrow$ FINDLR($E$, $T_{c_r}$, $\alpha$)
 9:         if $\lambda_\ell \neq \texttt{nil}$ and $\lambda_r \neq \texttt{nil}$ then
10:             return $(E, \texttt{nil}, \texttt{nil}) \cup \lambda_\ell \cup \lambda_r$
11:         return nil
12:     $f \leftarrow \alpha(r)$
13:     $(t_\ell, t_r) \leftarrow$ BINARYSEARCH($E$, $T$, $\alpha$, $c_\ell$, $c_r$, $f$)
14:     $\lambda_\ell \leftarrow$ FINDLR($E[f \leq t_\ell + 1]$, $T_{c_\ell}$, $\alpha$)
15:     $\lambda_r \leftarrow$ FINDLR($E[f > t_r - 1]$, $T_{c_r}$, $\alpha$)
16:     return $(E, t_\ell, t_r) \cup \lambda_\ell \cup \lambda_r$
```

Now we want to show that the construction of Π is correct, that is:

Claim 2. *$\Pi(v)$ is a pool set for v for every node $v \in V(T)$.*

We start proving (a): let v be a negative leaf and suppose there is an element $e^+ \in E^+ \cap E_v$, then there is no threshold assignment for T that would correctly classify e. The correctness for (b) is obtained by a symmetric argument. Let us consider (c) and let $f \in S$ be the feature at v: thanks to the preprocessing, for every $e_\ell \in \Pi(c_\ell)$ and $e_r \in \Pi(c_r)$, we know that $f(e_\ell) < f(e_r)$. This means that either every element of $\Pi(c_\ell)$ is sent to c_ℓ or every element of $\Pi(c_r)$ is sent to c_r: by induction the statement is proven. Finally we consider statement (d). If every element of $\Pi(c_\ell)$ is sent to c_ℓ or every element of $\Pi(c_r)$ is set to c_r, the statement is proven by induction. Otherwise, there is an example $e_\ell \in \Pi(c_\ell)$ that ends in c_r and an example $e_r \in \Pi(c_r)$ that ends in c_ℓ. This means v is an unknown feature node in T_t that distinguishes two examples in $\Pi(v)$, namely e_ℓ and e_r. This proves Claim 2.

In particular, let us consider the pool set $\Pi(r)$ for the root r of T, we define $\Pi(T) := \Pi(r)$. In this way given T, we are able to compute the corresponding pool set.

Let S be a support set for a CI E, we stay that $B \subseteq feat(E)$ is a *branching set* for S if for every minimal DT T for E such that $S \subset feat(T)$ then $B \cap (feat(T) \setminus S) \neq \emptyset$.

▶ **Lemma 25.** *There is a $\mathcal{O}(2^{d^2/2} s^{2s+1} n^{1+o(1)} \log n)$ time algorithm that given a support set S computes a branching set R_0 for S of size at most $s^{2s+3}\delta_{\max}$.*

Proof. Let E be a CI, a support set S for E and an integer s. We start by enumerating all partial DTs of size at most s such that the set of supported features is exactly S. For every such partial DT T, thanks to Lemma 24, we are able to obtain the set E_v for every node $v \in V(T)$ in time $\mathcal{O}(2^{d^2/2} n^{1+o(1)} \log n)$. In a leaves-to-root fashion, we are able to compute the set $\Pi(v)$ for every node $v \in V(T)$ and ultimately $\Pi(T)$.

Let $R(T)$ be the set of all the features in $feat(E) \setminus S$ that distinguish at least two examples in $\Pi(T)$. The algorithm returns the set of features R_0 obtained by considering the union of the sets $R(T)$ over all these partial DTs T for S of size at most s. By Lemma 1 this algorithm runs in time $\mathcal{O}(2^{d^2/2} s^{2s+1} n^{1+o(1)} \log n)$.

Algorithm 2 Algorithm to compute the pair $(th_L(r), th_R(r))$ for the root r of T

Input: CI E, pseudo DT T, S-feature assignment α for T, feature f of the root of T, left child c_ℓ of the root of T, right child c_r of the root of T

Output: maximum threshold t_ℓ in $D_E(f)$ for f such that (T_{c_ℓ}, α) can be extended to a DT for $E[f \le t_\ell]$ and minimum threshold t_r in $D_E(f)$ for f such that (T_{c_r}, α) can be extended to a DT for $E[f > t_r]$

```
1: function binarySearch($E$, $T$, $\alpha$, $f$, $c_\ell$, $c_r$)
2:     $D \leftarrow$ "array containing all elements in $D_E(f)$ in
           ascending order"
3:     $L \leftarrow 0$; $R \leftarrow |D_E(f)| - 1$; $b \leftarrow 0$
4:     while $L \le R$ do
5:         $m \leftarrow \lfloor (L+R)/2 \rfloor$
6:         if FINDTH($E[f \le D[m]]$, $T_{c_\ell}$, $\alpha$) = TRUE then
7:             $L \leftarrow m + 1$; $b \leftarrow 1$
8:         else
9:             $R \leftarrow m - 1$; $b \leftarrow 0$
10:    if $b = 1$ then
11:        $t_\ell \leftarrow D[m]$
12:    $t_\ell \leftarrow D[m-1]$                ▷ assuming that $D[-1] = D[0] - 1$
13:    $L \leftarrow 0$; $R \leftarrow |D_E(f)| - 1$; $b \leftarrow 0$
14:    while $L \le R$ do
15:        $m \leftarrow \lfloor (L+R)/2 \rfloor$
16:        if FINDTH($E[f > D[m]]$, $T_{c_r}$, $\alpha$) = TRUE then
17:            $R \leftarrow m - 1$; $b \leftarrow 1$
18:        else
19:            $L \leftarrow m + 1$; $b \leftarrow 0$
20:    if $b = 1$ then
21:        $t_r \leftarrow D[m]$
22:    $t_r \leftarrow D[m+1]$        ▷ assuming that $D[|D_E(f)|] = D[|D_E(f)| - 1] + 1$
23:    return $(t_r, t_r)$
```

Now we show the size of R_0 is bounded. By construction $|\Pi(T)| \le |T| \le s$; for every two distinct elements of $\Pi(T)$, by definition, there are at most $\delta_{\max}$ features that distinguish such two examples. This means that $|R(T)| \le s^2\delta_{\max}$ and so R_0 has size at most $s^{2s+3}\delta_{\max}$.

We are left to show that R_0 is a branching set for S. Let T be a minimal DT for E such that $S \subset feat(T)$ and suppose by contradiction that $R_0 \cap (feat(T) \setminus S = \emptyset$. In particular we have that $R(T) \cap (feat(T) \setminus S = \emptyset$. This means that every unknown feature of T does not distinguish any two elements in $\Pi(T)$. By Claim 2, $\Pi(T) = \Pi(r)$, where r is the root of T, is a pool set and so T can not be extended to a DT for E, which is a contradiction. ◀

▶ **Lemma 26** ([23]). *Let E be a CI and let k be an integer. Then there is an algorithm that in time $\mathcal{O}(\delta_{\max}(E)^k|E|)$ enumerates all (of the at most $\delta_{\max}(E)^k$) minimal support sets of size at most k for E.*

▶ **Lemma 27** ([23]). *Let T be a DT of minimum size for E and let S be a support set contained in feat(T). Then, the set $R = $ feat$(T) \setminus S$ is useful.*

▶ **Theorem 28.** Minimum Decision Tree Size *is fixed-parameter tractable prarametrized by $\delta_{\max} + s$.*

Proof. We start by presenting the algorithm for Minimum Decision Tree Size, which is illustrated in Algorithm 4 and Algorithm 5.

Algorithm 3 Algorithm to compute the threshold assignment for a pseudo DT and a feature assignment.

```
Input: CI $E$, pseudo DT $T$, $S$-feature assignment $\alpha$ for $T$
Output: TRUE if all examples in $E$ can be correctly classified by $(T, \alpha)$, FALSE otherwise
 1: function findTH($E$, $T$, $\alpha$)
 2:     $r \leftarrow$ “root of $T$”
 3:     if $r$ is a leaf then
 4:         if $E$ is not uniform then
 5:             return FALSE
 6:         return TRUE
 7:     $c_\ell, c_r \leftarrow$ “left child and right child of $r$”
 8:     if $r$ has an unknown feature then
 9:         $\lambda_\ell \leftarrow$ findTH($E$, $T_{c_\ell}$, $\alpha$)
10:         $\lambda_r \leftarrow$ findTH($E$, $T_{c_r}$, $\alpha$)
11:         if $\lambda_\ell =$ TRUE and $\lambda_r =$ TRUE then
12:             return TRUE
13:         return FALSE
14:     $f \leftarrow \alpha(r)$
15:     $t \leftarrow$ BinarySearch($E$, $T$, $\alpha$, $c_\ell$, $f$)
16:     $\lambda_r \leftarrow$ findTH($E[f > t]$, $T_{c_r}$, $\alpha$)
17:     if $\lambda_r =$ FALSE then
18:         return FALSE
19:     return TRUE
```

Given a CI E and an integer s, the algorithm returns a DT of minimum size among all DTs of size at most s if such a DT exists and otherwise the algorithm returns `nil`. The algorithm **minDT** starts by computing the set $\mathcal{S}$ of all minimal support sets for E of size at most s, which because of Lemma 26 results in a set $\mathcal{S}$ of size at most (). In Line 4 the algorithm then interates over all sets S in $\mathcal{S}$ and calls the function **minDTS** given in Algorithm 5 for E, s, and S, which returns a DT of minimum size among all DTs T for E of size at most s such that $S \subseteq feat(T)$. It then updates the currently best decision tree B if necessary with the DT found by the function **minDTS**. Moreover, if the best DT found after going through all sets in $\mathcal{S}$ has size at most s, it is returned (in Line 9), otherwise the algorithm returns `nil`. Finally, the function **minDTS** given in Algorithm 5 does the following. It first computes a DT T of minimum size that uses exactly the features in S using Lemma 23. It then tries to improve upon T with the help of useful sets. That is, it uses Lemma 25 to compute the branching set R_0. It then interates over all (of the at most ()) features $f \in R_0$ (using the for-loop in Line 4), and calls itself recursively on the feature set $S \cup \{f\}$. If this call finds a smaller DT, then the current best DT B is updated. Finally, after the for-loop the algorithm either returns B if its size is less then s or `nil` otherwise.

Towards showing the correctness of Algorithm 4, consider the case that E has a DT of size at most s and let T be a such a DT of minimum size. Because of Observation **??**, $feat(T)$ is a support set for E and therefore $feat(T)$ contains a minimal support set S of size at most s. Because the for-loop in Line 4 of Algorithm 4 interates over all minimal support sets of size at most s for E, it follows that Algorithm 5 is called with parameters E, s, and S. If $feat(T) = S$, then B is set to a DT for E of size $|T|$ in Line 2 of Algorithm 5 and the algorithm will output a DT of size at most $|T|$ for E. If, on the other hand, $feat(T) \setminus S \neq \emptyset$, then because T has minimum size and S is a support set for E with $S \subseteq feat(T)$, we obtain from Lemma 27 that the set $R = feat(T) \setminus S$ is useful for S. Therefore, because of Lemma 25, R has to contain a feature f from the set R_0 computed in Line 3. It follows that Algorithm 5 is called with parameters E, s, and $S \cup \{v\}$. From now onwards the argument repeats and since $R_0 \neq \emptyset$ the process stops after at most $s - |S|$ recursive calls after which a DT for E of size at most $|T|$ will be computed in Line 2 of Algorithm 5. Finally, it is easy to see that if Algorithm 4 outputs a DT T, then it is a valid solution. This is because, T must have been computed in Line 2 of Algorithm 5, which implies that T is a DT for E. Moreover, T has size at most s, because of Line 8 in Algorithm 4.

To analyse the run-time of the algorithm, we first remark that the whole algorithm can be seen as a bounded-depth search tree algorithm, i.e., a branching algorithm with small recursion depth and few branches at every node. In particular, every recursive call adds at least one feature to the set of features bounding the recursion depth to at most s. Moreover, every feature that is added is either added in Line 2 of Algorithm 4, when enumerating all minimal support sets, in which case there are at most $\delta_{\max}(E)$ branches or the feature is added in Line 5 of Algorithm 5, in which case there are at most $|R_0| \leq s^{2s+3}\delta_{\max}(E)$ branches. It follows that the algorithm can be seen as a branching algorithm of depth at most s with at most $s^{2s+3}\delta_{\max}(E) = \max\{s^{2s+3}\delta_{\max}(E), \delta_{\max}(E)\}$ branches at every step. Therefore, the total run-time of the algorithm is at most the number of nodes in the branching tree, i.e., at most $(s^{2s+3}\delta_{\max}(E))^s$, times the maximum time required in one recursive call. Now the maximum time required for one recursive call is dominated by the time spend in Line 2 of Algorithm 5, i.e., the time required to compute a DT of minimum size using exactly the features in S with the help of Theorem 22, which is at most $2^{\mathcal{O}(s^2)}\|E\|^{1+o(1)} \log \|E\|$. Therefore, we obtain $(s^{2s+3}\delta_{\max}(E))^s 2^{\mathcal{O}(s^2)}\|E\|^{1+o(1)} \log \|E\|$ as the total run-time of the algorithm, which shows that DTS is fixed-parameter tractable parameterized by $s + \delta_{\max}(E)$. ◀

Algorithm 4 Main method for finding a DT of minimum size.

Input: CI E and integer s
Output: DT for E of minimum size (among all DTs of size at most s) if such a DT exists, otherwise `nil`

```
 1: function minDT($E$, $s$)
 2:     $\mathcal{S} \leftarrow$ "set of all minimal support sets for $E$ of size at most $s$ using Lemma 26"
 3:     $B \leftarrow$ nil
 4:     for $S \in \mathcal{S}$ do
 5:         $T \leftarrow$ minDTS($E$, $s$, $S$)
 6:         if ($T \neq$ nil) and ($B =$ nil or $|B| > |T|$) then
 7:             $B \leftarrow T$
 8:     if $B \neq$ nil and $|B| \leq s$ then
 9:         return $B$
10:     return nil
```

5 Conclusion

We have initiated the study of the parameterized complexity of learning DTs from data. Our main tractability result provides novel insights into the structure of DTs and is based on the NLC-width parameter that seems to be well suited to measure the complexity of input instances for the problem.

The problem of learning DTs comes in many variants and flavors, which opens up a wide range of new research directions to explore. For instance:

Algorithm 5 Method for finding a DT of minimum size using at least the features in a given support set S.

```
Input: CI $E$, integer $s$, support set $S$ for $E$ with $|S| \leq s$
Output: DT of minimum size among all DTs $T$ for $E$ of size at most $s$ such that $S \subseteq feat(T)$; if no
    such DT exists, nil
 1: function minDTS($E$, $s$, $S$)
 2:     $B \leftarrow$ “compute a DT of minimum size for $E$ using exactly the features in $S$ using Theorem 23”
 3:     $R_0 \leftarrow$ “compute the branching set $R_0$ for $S$ using Lemma 25”
 4:     for $f \in R_0$ do
 5:         $T \leftarrow$ minDTS($E$, $s$, $S \cup \{f\}$)
 6:         if $T \neq$ nil and $|T| < |B|$ then
 7:             $B \leftarrow T$
 8:     if $|B| \leq s$ then
 9:         return $B$
10:     return nil
```

- What other (structural) parameters can be exploited to efficiently learn DTs? Is learning DTs of small size fixed-parameter tractable parameterized by the rank-width of $G_I(E)$?
- Instead of learning DTs of small size, one often wants to learn DTs of small height. Therefore, it is natural to ask whether our approach can be also used in this setting. While one can adapt our approach to obtain an XP-algorithm for learning DTs of small height parameterized by NLC-width, it is not clear to us whether the problem also allows for an fpt-algorithm.
- Can we extend our approach to CIs, where features range over an arbitrary domain? In this case, one usually still uses DTs that make binary decisions (i.e. whether a feature is smaller equal or larger than a given threshold). While it is relatively easy to see that our approach can be extended if the domain's size (for every feature) is bounded or used as an additional parameter, it is not clear what happens if the size of the domain is allowed to grow arbitrarily.

References

1 Jørgen Bang-Jensen and Gregory Gutin. *Digraphs*. Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, second edition, 2009.

2 Christian Bessiere, Emmanuel Hebrard, and Barry O'Sullivan. Minimising decision tree size as combinatorial optimisation. In Ian P. Gent, editor, *Principles and Practice of Constraint Programming - CP 2009*, pages 173–187, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

3 Hans L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM J. Comput.*, 25(6):1305–1317, 1996.

4 Endre Boros, Yves Crama, Peter L. Hammer, Toshihide Ibaraki, Alexander Kogan, and Kazuhisa Makino. Logical analysis of data: classification with justification. *Ann. Oper. Res.*, 188(1):33–61, 2011.

5 Endre Boros, Vladimir Gurvich, Peter L. Hammer, Toshihide Ibaraki, and Alexander Kogan. Decomposability of partially defined Boolean functions. *Discr. Appl. Math.*, 62(1-3):51–75, 1995.

6 Endre Boros, Takashi Horiyama, Toshihide Ibaraki, Kazuhisa Makino, and Mutsunori Yagiura. Finding essential attributes from binary data. *Ann. Math. Artif. Intell.*, 39:223–257, 11 2003. doi:10.1023/A:1024653703689.

7 Endre Boros, Toshihide Ibaraki, and Kazuhisa Makino. Variations on extending partially defined Boolean functions with missing bits. *Information and Computation*, 180(1):53–70, 2003.

8 Yves Crama, Peter L. Hammer, and Toshihide Ibaraki. Cause-effect relationships and partially defined Boolean function. *Ann. Oper. Res.*, 16:299–326, 1988.

9 Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms.* Springer, 2015. doi:10.1007/978-3-319-21275-3.

10 Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications.* IOS Press, 2020.

11 Reinhard Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics.* Springer Verlag, New York, 2nd edition, 2000.

12 Finale Doshi-Velez and Been Kim. A roadmap for a rigorous science of interpretability. *CoRR*, abs/1702.08608, 2017. URL: http://arxiv.org/abs/1702.08608, arXiv:1702.08608.

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

14 Wolfgang Espelage, Frank Gurski, and Egon Wanke. Deciding clique-width for graphs of bounded tree-width. *J. Graph Algorithms Appl.*, 7(2):141–180, 2003.

15 Bryce Goodman and Seth R. Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3):50–57, 2017.

16 Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976.

17 Toshihide Ibaraki, Yves Crama, and Peter L. Hammer. *Partially defined Boolean functions*, page 511–563. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2011.

18 Daniel T. Larose. *Discovering knowledge in data.* Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2005. An introduction to data mining.

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

20 E.J. McCluskey. *Introduction to the Theory of Switching Circuits.* Electrical and electronic engineering series. Princeton University series. McGraw-Hill, 1965.

21 Don Monroe. AI, explain yourself. *AI Communications*, 61(11):11–13, 2018. doi:10.1145/3276742.

22 Sreerama K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Min. Knowl. Discov.*, 2(4):345–389, 1998. doi:10.1023/A:1009744630224.

23 Sebastian Ordyniak and Stefan Szeider. Parameterized complexity of small decision tree learning. In *Proceedings of AAAI'21, the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 6454–6462. AAAI Press, 2021.

24 Sang-il Oum. Approximating rank-width and clique-width quickly. *ACM Transactions on Algorithms*, 5(1), 2008.

25 J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. doi:10.1023/A:1022643204877.

26 Felix Reidl, Peter Rossmanith, Fernando Sánchez Villaamil, and Somnath Sikdar. A faster parameterized algorithm for treedepth. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, volume 8572 of *Lecture Notes in Computer Science*, pages 931–942. Springer, 2014.

27 Richard Stanley and Eric W. Weisstein. Catalan number, from mathworld–a wolfram web resource, 2015.

28 Egon Wanke. k-NLC graphs and polynomial algorithms. *Discr. Appl. Math.*, 54(2-3):251–266, 1994. Efficient algorithms and partial k-trees.