Journal of Machine Learning Research 24 (2023) 1-58 Submitted 4/21; Revised 10/22; Published 3/23

On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results

Marcelo Arenas MARENAS@ING.PUC.CL
(a) Department of Computer Science & Institute for Mathematical and Computational Engineering, School of Engineering & Faculty of Mathematics, Universidad Católica de Chile
(b) IMFD Chile

Pablo Barceló PBARCELO@UC.CL
(a) Institute for Mathematical and Computational Engineering, School of Engineering & Faculty of Mathematics, Universidad Católica de Chile
(b) IMFD Chile (c) National Center for Artificial Intelligence, CENIA Chile

Leopoldo Bertossi LEOPOLDO.BERTOSSI@SKEMA.EDU
SKEMA Business School, Montreal, Canada

Mikaël Monet MIKAEL.MONET@INRIA.FR
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

Editor: David Jensen

Abstract

Scores based on Shapley values are widely used for providing explanations to classification results over machine learning models. A prime example of this is the influential SHAP-score, a version of the Shapley value that can help explain the result of a learned model on a specific entity by assigning a score to every feature. While in general computing Shapley values is a computationally intractable problem, we prove a strong positive result stating that the SHAP-score can be computed in polynomial time over *deterministic and decomposable Boolean circuits* under the so-called product distributions on entities. Such circuits are studied in the field of Knowledge Compilation and generalize a wide range of Boolean circuits and binary decision diagrams classes, including binary decision trees, Ordered Binary Decision Diagrams (OBDDs) and Free Binary Decision Diagrams (FBDDs). Our positive result extends even beyond binary classifiers, as it continues to hold if each feature is associated with a finite domain of possible values.

We also establish the computational limits of the notion of SHAP-score by observing that, under a mild condition, computing it over a class of Boolean models is always polynomially as hard as the model counting problem for that class. This implies that both determinism and decomposability are essential properties for the circuits that we consider, as removing one or the other renders the problem of computing the SHAP-score intractable (namely, #P-hard). It also implies that computing SHAP-scores is #P-hard even over the class of propositional formulas in DNF. Based on this negative result, we look for the existence of fully-polynomial randomized approximation schemes (FPRAS) for computing SHAP-scores over such class. In stark contrast to the model counting problem for DNF

formulas, which admits an FPRAS, we prove that no such FPRAS exists (under widely believed complexity assumptions) for the computation of SHAP-scores. Surprisingly, this negative result holds even for the class of monotone formulas in DNF. These techniques can be further extended to prove another strong negative result: Under widely believed complexity assumptions, there is no polynomial-time algorithm that checks, given a monotone DNF formula φ and features x, y, whether the SHAP-score of x in φ is smaller than the SHAP-score of y in φ.

Keywords: Explainable AI, Shapley values, SHAP score, knowledge compilation, FPRAS

1. Introduction

Context. Explainable artificial intelligence has become an active area of research. Central to it is the observation that artificial intelligence (AI) and machine learning (ML) models cannot always be blindly applied without being able to interpret or explain their results. For example, someone who applies for a loan and sees the application rejected by an algorithmic decision-making system would like to receive from the system an explanation for this decision. In ML, explanations have been commonly considered for classification algorithms, and there are different approaches. In particular, explanations can be *global* – focusing on the general input/output relation of the model –, or *local* – focusing on how individual features affect the decision of the model for a specific input, as in the loan example above (Ribeiro et al., 2016; Lundberg and Lee, 2017; Chen et al., 2018; Bertossi et al., 2020). Recent literature has strengthened the importance of the latter by showing their ability to provide explanations that are often overlooked by global explanations (Molnar, 2020). In this work we concentrate on local explanations.

One way to define local explanations is by considering feature values as players in a coalition game that jointly contribute to the outcome. More concretely, one treats a feature value's contribution from the viewpoint of game theory, and ties it to the question of how to properly distribute wealth (profit) among collaborating players. This problem was approached in general terms in Shapley (1953). One can use the established concepts and techniques that he introduced in the context of cooperative game theory; and, more specifically, use the popular *Shapley value* as a measure of the contribution of a player to the common wealth associated with a multi-player game. It is well known that the Shapley value possesses properties that cast it as natural and intuitive. Actually, the Shapley value emerges as the only function that enjoys those desirable properties (Roth, 1988). The Shapley value has been widely applied in different disciplines, in particular in computer science (Hunter and Konieczny, 2010; Michalak et al., 2013; Cesari et al., 2018; Livshits et al., 2020, 2021); and in machine learning it has been applied to the explanation of classification results, in its incarnation as the SHAP-*score* (Lundberg and Lee, 2017; Lundberg et al., 2020). Here, the players are the feature values of an entity under classification.

In this paper, we concentrate on the SHAP-*score* for classification models. It has a clear, intuitive, combinatorial meaning, and inherits all the good properties of the Shapley value. Accordingly, an explanation for a classification result takes the form of a set of feature values that have a high, hopefully maximum, SHAP-*score*. We remark that SHAP-scores have attracted the attention of the ML community and have found several applications and extensions (Rathi, 2019; Fidel et al., 2020; Bertossi et al., 2020; Merrick and Taly, 2020;

Takeishi and Kawahara, 2020; Covert and Lee, 2021; Kumar et al., 2020). However, its fundamental and computational properties have not been investigated much.

Problems studied in the paper. For a given classification model M, entity $\mathbf{e}$ and feature x, the SHAP-score $\mathsf{SHAP}(M, \mathbf{e}, x)$ intuitively represents the importance of the feature value $\mathbf{e}(x)$ to the classification result $M(\mathbf{e})$. In its general formulation, $\mathsf{SHAP}(M, \mathbf{e}, x)$ is a weighted average of differences of expected values of the outcomes (c.f. Section 2 for its formal definition). Unfortunately, computing quantities that are based on the notion of Shapley value is in general intractable. Indeed, in many scenarios the computation turns out to be #P-hard (Faigle and Kern, 1992; Deng and Papadimitriou, 1994; Livshits et al., 2021; Bertossi et al., 2020), which makes the notion difficult to use – if not impossible – for practical purposes (Arora and Barak, 2009). Therefore, natural questions are: "For what kinds of classification models the computation of the SHAP-score can be done efficiently?", "Can one obtain lower computational complexity if one has access to the internals of the classification model?", or again "In cases where exact computation is intractable, can we efficiently approximate the SHAP-score?".

In Lundberg et al. (2020) the claim is made that for certain models based on decision trees the computation of the SHAP-score is tractable. In this work we go deeper into these results, in particular, formulating and establishing them in precise terms, and extending them for a larger class of classification models. We also identify classes of models for which SHAP-score computation is intractable. In such cases, we investigate the problem of existence and computation of a good approximation in the form of a *fully polynomial-time randomized approximation scheme* (FPRAS). Recall that FPRAS are tractable procedures that return an answer that is, with high probability, close to the correct answer.

Given the high computational complexity of the SHAP-score, one might try to solve related problems, other than the exact and approximate computation of all the features' scores, that could still be useful in practice. For instance, we consider the problem that consists in deciding, for a pair of feature values, which of the two has the highest score. This problem could indeed be used to compute a ranking of (all or the highest) SHAP-scores, without computing them explicitly. We also address this problem in this paper.

Model studied in the paper. We focus mainly on binary classifiers with *binary feature values* (i.e., propositional features that can take the values 0 or 1), and that return 1 (accept) or 0 (reject) for each entity. We will call these *Boolean classifiers*. The restriction to binary inputs can be relevant in many practical scenarios where the features are of a propositional nature. Still, we consider classifiers with possibly non-binary features, but binary outcomes, in Section 4. The second assumption that we make is that the underlying probability distributions on the population of entities are what we call *product distributions*, where each binary feature x has a probability $\mathrm{p}(x)$ of being equal to 1, independently of the other feature values. This includes, as a special case, the *uniform probability distribution* when each $\mathrm{p}(x)$ is $\frac{1}{2}$. Product distributions are also known as *fully-factorized* in the literature (Van den Broeck et al., 2021, 2022). They have received considerable attention in the context of computing score-based explanations, as they combine good computational properties with enough flexibility to model relevant practical scenarios (Strumbelj and Kononenko, 2010; Datta et al., 2016; Lundberg and Lee, 2017).

Positive results on the complexity of computation of SHAP-scores in the paper are obtained for Boolean classifiers defined as *deterministic and decomposable Boolean circuits*. This is a widely studied model in *knowledge compilation* (Darwiche, 2001; Darwiche and Marquis, 2002). Such circuits encompass a wide range of Boolean circuits and binary decision diagrams classes that are considered in knowledge compilation, and more generally in AI. For instance, they generalize *binary decision trees*, *ordered binary decision diagrams* (OBDDs), *free binary decision diagrams* (FBDDs), and *deterministic and decomposable negation normal norms* (d-DNNFs) (Darwiche, 2001; Amarilli et al., 2020; Darwiche and Hirth, 2020). These circuits are also known under the name of *tractable Boolean circuits*, that is used in recent literature (Shih et al., 2019a; Shi et al., 2020; Shih et al., 2018a,b, 2019b; Peharz et al., 2020). Readers who are not familiar with knowledge compilation can simply think about deterministic and decomposable circuits as a tool for analyzing in a uniform manner the computational complexity of the SHAP-score on several Boolean classifier classes.

In turn, our negative results on the complexity of computation of SHAP-scores in the paper are obtained over the class of propositional formulas in DNF. In addition to being a well-known restriction of the class of propositional formulas for which the satisfiability problem is tractable, DNF formulas define an extension of deterministic and decomposable Boolean circuits. In fact, DNF formulas can be seen as decomposable, although not necessarily deterministic, Boolean circuits.

Our results. Our main contributions are the following.

1. *Tractability for a large class of Boolean classifiers.* We provide a polynomial time algorithm that computes the SHAP-score for deterministic and decomposable Boolean circuits under product distributions over the entity population (Theorem 2). We obtain as a corollary that the SHAP-score for Boolean classifiers given as binary decision trees, OBDDs, FBDDs and d-DNNFs can be computed in polynomial time.

2. *Tractability for non-binary classifiers.* We extend the aforementioned tractability result to the case of classifiers with non-binary features, which take the form of *non-binary Boolean circuits* (Theorem 5). In these circuits, the nodes may now contain equalities of the form $x = v$, where x is a feature, and v is a value in the domain of x. The outcome of the classifier is still binary.

3. *Limits of tractability.* We observe that, under a mild condition, computing the SHAP-score on Boolean classifiers in a class is always polynomially as hard as the *model counting* problem for that class (Lemma 7). This leads to intractability for the problem of computing the SHAP-score for all classes of Boolean classifiers for which model counting is intractable. An important example of this corresponds to the class of propositional formulas in DNF. As a corollary, we obtain that each one of the *determinism* assumption and the *decomposability* assumption is necessary for tractability (Theorem 6). These results even hold for the uniform distribution.

4. *Non-approximability for DNF formulas.* We give a simple proof that, under widely believed complexity assumptions, there is no FPRAS for the computation of the SHAP-score with Boolean classifiers represented as DNF formulas (Proposition 8). This

holds even under the uniform distribution. This result establishes a stark contrast with the model counting problem for DNF formulas, which admits an FPRAS (Karp et al., 1989). We further strengthen this by showing that the non-approximability result even holds for Boolean formulas represented as 2-POS-DNF, i.e., with every conjunct containing at most two positive literals (and no negative literal), and for the uniform distribution (Theorem 9). The proof of this last result is quite involved and is based on a non-approximability result in relation to the size of cliques in graphs (Feige et al., 1996; Arora and Safra, 1998; Arora et al., 1998).

5. *Impossibility of comparing* SHAP*-scores for DNF formulas.* Consider the problem of verifying, given a DNF formula φ and features x, y, whether the SHAP-score of x in φ is smaller than the SHAP-score of y in φ. Under widely believed complexity assumptions, we establish that this problem of comparing two SHAP-scores cannot be approximated in polynomial time, even for the case of monotone DNF formulas (Theorem 16).

Related work. Our contributions should be compared to the results obtained in contemporaneous papers by Van den Broeck et al. (Van den Broeck et al., 2021, 2022). There, the authors establish the following theorem: for every class $\mathcal{C}$ of classifiers and under product distributions, the problem of computing the SHAP-score for $\mathcal{C}$ is polynomial-time equivalent to the problem of computing the expected value for the models in $\mathcal{C}$ (at least under mild assumptions on $\mathcal{C}$). Since computing expectations is in polynomial time for tractable Boolean circuits, this in particular implies that computing the SHAP-score is in polynomial time for the circuits that we consider; in other words, their results capture our main positive result. However, there is a fundamental difference in the approach taken to show tractability: their reduction uses multiple oracle calls to the problem of computing expectations, whereas we provide a more direct algorithm to compute the SHAP-score on these circuits.

The exact computation or approximation of the Shapley value applied to database tuples is investigated in Livshits et al. (2021); Deutch et al. (2022). In Bertossi et al. (2020), approximations of the SHAP-score are used for experimental purposes and comparisons with other scores, such as RESP and Rudin's FICO-score (Chen et al., 2018). However, an empirical distribution is used for the approximate computation of the SHAP-score, which leads to a simple, non-probabilistic weighted average, and to the restriction of the counterfactual versions of an entity to those in the available sample. In Bertossi and Leon (2023), the algorithm presented in our work (c.f. Section 3.1.3) has been used to efficiently compute SHAP-scores, under the uniform distribution, for outcomes from binary neural networks, after compiling the latter into deterministic and decomposable Boolean circuits.

Building on Livshits et al. (2021) and on the conference version of the current article, the authors of Deutch et al. (2022) use knowledge compilation techniques that are similar to ours to develop a polynomial-time algorithm that is able to compute a version of the Shapley value tailored to a database context. They moreover propose an algorithm for computing approximations of the Shapley values of tuples, but this algorithm does not come with any theoretical guarantees.

Last, regarding approximation, there is a large body of work that aims at approximating Shapley values using Monte Carlo techniques, see for instance Castro et al. (2009); Okhrati and Lipani (2021); Deutch et al. (2021); Mitchell et al. (2022). The results of such works

are upper bounds for diverse settings of Shapley values – i.e., considering different game functions – but we are not aware of such results for the specific game function (the SHAP-score) that we study here, for instance, theoretical results that would yield an FPRAS for the SHAP-score. Besides, these studies do not usually contain lower bounds, since their focus is on obtaining tractable approximations via Monte Carlo or other sampling schemes. By contrast, obtaining lower bounds is the topic of our Sections 6 and 7, where we show the non-existence of an FPRAS for approximating the SHAP-score of DNF formulas, as well as the non-membership in BPP of comparing SHAP-scores for such formulas.

This paper is a considerable extension of the conference paper (Arenas et al., 2021). In addition to containing full proofs of all the results of (Arenas et al., 2021), we present here several new results. Among them, we provide a detailed analysis of approximability of the SHAP-score, a complexity analysis of the problem of comparing SHAP-scores, and an extension of the results for deterministic and decomposable Boolean circuits to the case of non-binary features.

Paper structure. We give preliminaries in Section 2. In Section 3, we prove that the SHAP-score can be computed in polynomial time for deterministic and decomposable Boolean circuits under product probability distributions. We extend this tractability result in Section 4 to non-binary deterministic and decomposable Boolean circuits. In Section 5, we establish the computational limits of the exact computation of the SHAP-score, while Section 6 studies the (non-) approximability properties of this score. The problem of comparing the SHAP-scores of different features is studied in Section 7. We conclude and discuss future work in Section 8.

2. Preliminaries

2.1 Entities, distributions and classifiers

Let X be a finite set of *binary*[1] *features*, also called *variables*. An *entity* over X is a function $\mathbf{e} : X \to \{0,1\}$.[2] We denote by $\mathsf{ent}(X)$ the set of all entities over X. On this set, we consider probability distributions that we call *product distributions*, defined as follows. Let $\mathrm{p} : X \to [0,1]$ be a function that associates to every feature $x \in X$ a probability value $\mathrm{p}(x) \in [0,1]$. Then, the *product distribution generated by* p is the probability distribution Π_{p} over $\mathsf{ent}(X)$ such that, for every $\mathbf{e} \in \mathsf{ent}(X)$ we have

$$\Pi_{\mathrm{p}}(\mathbf{e}) \quad := \quad \Big(\prod_{\substack{x \in X \\ \mathbf{e}(x)=1}} \mathrm{p}(x) \Big) \cdot \Big(\prod_{\substack{x \in X \\ \mathbf{e}(x)=0}} (1 - \mathrm{p}(x)) \Big).$$

That is, Π_{p} is the product distribution that is determined by pre-specified marginal distributions, and that makes the features take values independently from each other. We denote by $\mathcal{U}$ the *uniform probability distribution*, i.e., for every entity $\mathbf{e} \in \mathsf{ent}(X)$, we have that $\mathcal{U}(\mathbf{e}) := \frac{1}{2^{|X|}}$. Note that the uniform distribution can be obtained as a special case of product distribution, with Π_{p} invoking $\mathrm{p}(x) := 1/2$ for every $x \in X$.

1. We will come back to this assumption in Section 4, where we will consider non-binary features.
2. Equivalently, one can see an entity as a vector of binary values, with each coordinate corresponding to a given feature. We will however always use the functional point of view as it simplifies the notation in our proofs.

A *Boolean classifier* M over X is a function $M : \mathsf{ent}(X) \to \{0, 1\}$ that maps every entity over X to 0 or 1. We say that M *accepts* an entity $\mathbf{e}$ when $M(\mathbf{e}) = 1$, and that it *rejects* it if $M(\mathbf{e}) = 0$. Since we consider $\mathsf{ent}(X)$ to be a probability space, M can be regarded as a random variable.

2.2 The SHAP-score over Boolean classifiers

Let $M : \mathsf{ent}(X) \to \{0, 1\}$ be a Boolean classifier over the set X of features. Given an entity $\mathbf{e}$ over X and a subset $S \subseteq X$ of features, we define the set $\mathsf{cw}(\mathbf{e}, S)$ of entities that are *consistent with* $\mathbf{e}$ *on* S as $\mathsf{cw}(\mathbf{e}, S) := \{\mathbf{e}' \in \mathsf{ent}(X) \mid \mathbf{e}'(x) = \mathbf{e}(x) \text{ for each } x \in S\}$. Then, given an entity $\mathbf{e} \in \mathsf{ent}(X)$, a probability distribution $\mathcal{D}$ over $\mathsf{ent}(X)$, and $S \subseteq X$, we define the *expected value of* M *over* $X \setminus S$ *with respect to* $\mathbf{e}$ *under* $\mathcal{D}$ as

$$\phi_{\mathcal{D}}(M, \mathbf{e}, S) \;:=\; \mathbb{E}_{\mathbf{e}' \sim \mathcal{D}}\big[M(\mathbf{e}') \mid \mathbf{e}' \;\in \mathsf{cw}(\mathbf{e}, S)\big].$$

In other words, $\phi_{\mathcal{D}}(M, \mathbf{e}, S)$ is the expected value of M, conditioned on the inputs to coincide with $\mathbf{e}$ over each feature in S. For instance, if we take $\mathcal{D}$ to be the uniform distribution $\mathcal{U}$ over $\mathsf{ent}(X)$, this expression simplifies to

$$\phi_{\mathcal{U}}(M, \mathbf{e}, S) = \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)} \frac{1}{2^{|X \setminus S|}} M(\mathbf{e}').$$

The function $\phi_{\mathcal{D}}$ is then used in the general formula of the Shapley value (Shapley, 1953; Roth, 1988) to obtain the SHAP-score for feature values in $\mathbf{e}$, as follows.

Definition 1 (Lundberg and Lee (2017)) *Given a Boolean classifier M over a set of features X, a probability distribution $\mathcal{D}$ on $\mathsf{ent}(X)$, an entity $\mathbf{e}$ over X, and a feature $x \in X$, the* SHAP score of feature x on $\mathbf{e}$ with respect to M under $\mathcal{D}$ *is defined as*

$$\mathsf{SHAP}_{\mathcal{D}}(M, \mathbf{e}, x) \;:=\; \sum_{S \subseteq X \setminus \{x\}} \frac{|S|!\,(|X| - |S| - 1)!}{|X|!} \Big(\phi_{\mathcal{D}}(M, \mathbf{e}, S \cup \{x\}) - \phi_{\mathcal{D}}(M, \mathbf{e}, S) \Big). \quad (1)$$

In Section 5, we will use another equivalent expression of the SHAP-score, that we introduce now. For a permutation $\pi : X \to \{1, \ldots, n\}$ and $x \in X$, let S^x_π denote the set of features that appear before x in π. Formally, $S^x_\pi := \{y \in X \mid \pi(y) < \pi(x)\}$. Then, letting $\Pi(X)$ be the set of all permutations $\pi : X \to \{1, \ldots, n\}$, observe that Equation (1) can be rewritten as

$$\mathsf{SHAP}_{\mathcal{D}}(M, \mathbf{e}, x) \;=\; \frac{1}{|X|!} \sum_{\pi \in \Pi(X)} \big(\phi_{\mathcal{D}}(M, \mathbf{e}, S^x_\pi \cup \{x\}) - \phi_{\mathcal{D}}(M, \mathbf{e}, S^x_\pi) \big). \quad (2)$$

Thus, $\mathsf{SHAP}_{\mathcal{D}}(M, \mathbf{e}, x)$ is a weighted average of the contribution of feature x on $\mathbf{e}$ to the classification result, i.e., of the differences between having it and not, under all possible permutations of the other feature values. Observe that, from this definition, a high positive value of $\mathsf{SHAP}_{\mathcal{D}}(M, \mathbf{e}, x)$ intuitively means that setting x to $\mathbf{e}(x)$ strongly leans the classifier towards acceptance, while a high negative value of $\mathsf{SHAP}_{\mathcal{D}}(M, \mathbf{e}, x)$ means that setting x to $\mathbf{e}(x)$ strongly leans the classifier towards rejection.

2.3 Deterministic and decomposable Boolean circuits

A Boolean circuit over a set of variables X is a directed acyclic graph C such that

(i) Every node without incoming edges is either a *variable gate* or a *constant gate*. A variable gate is labeled with a variable from X, and a constant gate is labeled with either 0 or 1;

(ii) Every node with incoming edges is a *logic gate*, and is labeled with a symbol $\wedge$, $\vee$ or $\neg$. If it is labeled with the symbol $\neg$, then it has exactly one incoming edge;[3]

(iii) Exactly one node does not have any outgoing edges, and this node is called the *output gate of* C.

Such a Boolean circuit C represents a Boolean classifier in the expected way – we assume the reader to be familiar with Boolean logic –, and we write $C(\mathbf{e})$ for the value in $\{0, 1\}$ of the output gate of C when we evaluate C over the entity $\mathbf{e}$. We consider the *size* $|C|$ of the circuit to be its number of edges. Observe that, thanks to condition (iii), the number of gates of C is at most its number of edges plus one.

Several restrictions of Boolean circuits with good computational properties have been studied. Let C be a Boolean circuit over a set of variables X and g a gate of C. The Boolean circuit C_g over X is defined by considering the subgraph of C induced by the set of gates g' in C for which there exists a path from g' to g in C. Notice that g is the output gate of C_g. Then, an $\vee$-gate g of C is said to be *deterministic* if for every pair g_1, g_2 of distinct input gates of g, the Boolean circuits C_{g_1} and C_{g_2} are disjoint in the sense that there is no entity $\mathbf{e}$ that is accepted by both C_{g_1} and C_{g_2} (that is, there is no entity $\mathbf{e} \in \mathsf{ent}(X)$ such that $C_{g_1}(\mathbf{e}) = C_{g_2}(\mathbf{e}) = 1$). The circuit C is called *deterministic* if every $\vee$-gate of C is deterministic. The set $\mathsf{var}(g)$ is defined as the set of variables $x \in X$ such that there exists a variable gate with label x in C_g. An $\wedge$-gate g of C is said to be *decomposable*, if for every pair g_1, g_2 of distinct input gates of g, we have that $\mathsf{var}(g_1) \cap \mathsf{var}(g_2) = \emptyset$. Then C is called *decomposable* if every $\wedge$-gate of C is decomposable.

Example 1 We want to classify papers submitted to a conference as rejected (Boolean value 0) or accepted (Boolean value 1). Papers are described by propositional features **fg**, **dtr**, **nf** and **na**, which stand for "follows guidelines", "deep theoretical result", "new framework" and "nice applications", respectively. The Boolean classifier for the papers is given by the Boolean circuit in Figure 1. The input of this circuit are the features **fg**, **dtr**, **nf** and **na**, each of which can take value either 0 or 1, depending on whether the feature is present (1) or absent (0). The nodes with labels $\neg$, $\vee$ or $\wedge$ are logic gates, and the associated Boolean value of each one of them depends on the logical connective represented by its label and the Boolean values of its inputs. The output value of the circuit is given by the top node in the figure.

The Boolean circuit in Figure 1 is decomposable, because for each $\wedge$-gate the sets of features of its inputs are pairwise disjoint. For instance, in the case of the top node in Figure 1, the left-hand side input has $\{\mathbf{fg}\}$ as its set of features, while its right-hand side input

3. Recall that the fan-in of a gate is the number of its input gates. In our definition of Boolean circuits, we allow unbounded fan-in $\wedge$- and $\vee$-gates.

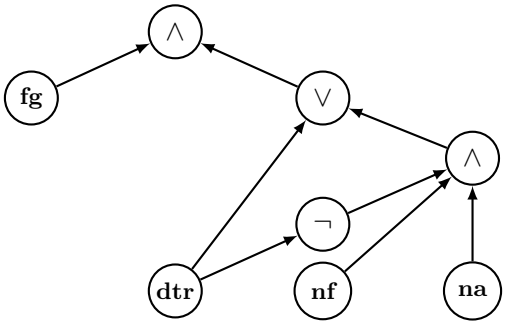


Figure 1: A deterministic and decomposable Boolean Circuit as a classifier.

has $\{\mathbf{dtr}, \mathbf{nf}, \mathbf{na}\}$ as its set of features, which are disjoint. Also, this circuit is deterministic as for every $\vee$-gate two (or more) of its inputs cannot be given value 1 by the same Boolean assignment for the features. For instance, in the case of the only $\vee$-gate in Figure 1, if a Boolean assignment for the features gives value 1 to its left-hand side input, then feature **dtr** has to be given value 1 and, thus, such an assignment gives value 0 to the right-hand side input of the $\vee$-gate. In the same way, it can be shown that if a Boolean assignment for the features gives value 1 to the right-hand side input of this $\vee$-gate, then it gives value 0 to its left-hand side input. □

We will use the fact that deterministic and decomposable Boolean circuits are closed under *conditioning*. Let C be a Boolean circuit over variables X, and let $x \in X$. We denote by C_{+x} (resp., C_{-x}) the Boolean circuit that is obtained from C by replacing every variable gate that is labeled with x by a constant 1-gate (resp, by a constant 0-gate). In the literature, C_{+x} (resp., C_{-x}) is called the conditioning by x (resp., by $\neg x$) of C. It is easy to check that, if C is deterministic and decomposable, then so are C_{+x} and C_{-x}.

As mentioned in the introduction, deterministic and decomposable Boolean circuits generalize many decision diagrams and Boolean circuits classes. We refer to (Darwiche, 2001; Amarilli et al., 2020) for detailed studies of knowledge compilation classes and of their precise relationships. For the reader's convenience, we explain in Appendix A how FBDDs and binary decision trees can be encoded in linear time as deterministic and decomposable Boolean circuits.

2.4 Complexity classes and encoding of probability values

In Section 5, we will consider the counting complexity class #P (Valiant, 1979) of problems that can be expressed as the number of accepting paths of a nondeterministic Turing machine running in polynomial time. Following Valiant (1979), we define #P-hardness using Turing reductions. Prototypical examples of #P-complete problems are counting the number of assignments that satisfy a propositional formula and counting the number of three-colorings of a graph. While it is known that FPTIME $\subseteq$ #P, where FPTIME is the class of functions that can be computed in polynomial time, this inclusion is widely believed to be strict. Therefore, proving that a problem is #P-hard implies, under such an assumption, that it cannot be solved in polynomial time.

In Sections 6 and 7 we will use the complexity classes RP and BPP. Recall that RP is the class of decision problems L for which there exists a polynomial-time probabilistic Turing Machine M such that: (a) if $x \in L$, then M accepts with probability at least $3/4$; and (b) if $x \notin L$, then M does not accept x. Moreover, BPP is defined exactly as RP but with condition (b) replaced by: (b') if $x \notin L$, then M accepts with probability at most $1/4$. Thus, BPP is defined as RP but allowing errors for both the elements that are and are not in L. Hence, PTIME $\subseteq$ RP $\subseteq$ BPP by definition. Besides, it is known that RP $\subseteq$ NP, and this inclusion is widely believed to be strict. Finally, it is not known whether BPP $\subseteq$ NP or NP $\subseteq$ BPP, but it is widely believed that NP is not included in BPP.

Finally, when considering problems where probabilities can be part of the input (such as in Theorem 2 or Theorem 5), it will always be implicit that such probabilities are given as rational numbers p/q for $p, q \in \mathbb{N}$, encoded as ordered pairs (p, q) where p and q are themselves encoded in binary.

3. Tractable Computation of the SHAP-Score

In this section we prove our main tractability result, namely, that computing the SHAP-score for Boolean classifiers given as deterministic and decomposable Boolean circuits can be done in polynomial time for product probability distributions.

Theorem 2 *Given as input a deterministic and decomposable circuit C over a set of features X, rational probability values $\mathrm{p}(x)$ for every feature $x \in X$, an entity $\mathbf{e} : X \to \{0, 1\}$, and a feature $x \in X$, the value $\mathsf{SHAP}_{\Pi_\mathrm{p}}(C, \mathbf{e}, x)$ can be computed in polynomial time.*

In particular, since binary decision trees, OBDDs, FBDDs and d-DNNFs are all restricted types of deterministic and decomposable circuits, we obtain as a consequence of Theorem 2 that this problem is also in polynomial time for these classes. For instance, for binary decision trees we obtain the following result.

Corollary 3 *Given as input a binary decision tree T over a set of features X, rational probability values $\mathrm{p}(x)$ for every feature $x \in X$, an entity $\mathbf{e} : X \to \{0, 1\}$, and a feature $x \in X$, the value $\mathsf{SHAP}_{\Pi_\mathrm{p}}(T, \mathbf{e}, x)$ can be computed in polynomial time.*

The authors of (Lundberg et al., 2020) provide an algorithm for computing the SHAP-score in polynomial time for decision trees, but, unfortunately, as stated in (Van den Broeck et al., 2021), this algorithm is not correct. Moreover, it is important to notice that Theorem 2 is a nontrivial extension of the result for decision trees, as it is known that deterministic and decomposable circuits can be exponentially more succinct than binary decision trees (in fact, than FBDDs) at representing Boolean classifiers (Darwiche, 2001; Amarilli et al., 2020).

We prove Theorem 2 in the remaining of this section. First, we prove in Section 3.1 that the problem can be solved in polynomial time, and we extract from this proof a first version of our algorithm. Then, in Section 3.2, we provide an optimized version of this algorithm, and argue why the proposed optimization reduces the complexity of the algorithm.

3.1 Polynomial time computability proof

In this section we prove that the SHAP-score can be computed in polynomial time. For a Boolean classifier M over a set of variables X, probability distribution $\mathcal{D}$ over $\mathsf{ent}(X)$, entity $\mathbf{e} \in \mathsf{ent}(X)$, and natural number $k \leq |X|$, we define the quantity

$$\mathsf{H}_{\mathcal{D}}(M, \mathbf{e}, k) \ := \sum_{\substack{S \subseteq X \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \mathcal{D}}[M(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)].$$

Our proof of Theorem 2 is divided into two modular parts. The first part, which is developed in Section 3.1.1, consists in showing that the problem of computing $\mathsf{SHAP}_{\Pi.}(\cdot,\cdot,\cdot)$ can be reduced in polynomial time to that of computing $\mathsf{H}_{\Pi.}(\cdot,\cdot,\cdot)$. This part of the proof is a sequence of formula manipulations, and it only uses the fact that deterministic and decomposable circuits are closed under conditioning on a variable value (recall the definition of conditioning from Section 2.3). In the second part of the proof, which is developed in Section 3.1.2, we show that computing $\mathsf{H}_{\Pi.}(\cdot,\cdot,\cdot)$ can be done in polynomial time for deterministic and decomposable Boolean circuits. It is in this part that the properties of deterministic and decomposable circuits are used. Finally, we extract Algorithm 1 from this proof in Section 3.1.3.

3.1.1 Reducing in polynomial-time from $\mathsf{SHAP}_{\Pi.}(\cdot,\cdot,\cdot)$ to $\mathsf{H}_{\Pi.}(\cdot,\cdot,\cdot)$

In this section we show that for deterministic and decomposable Boolean circuits, and under product distributions, the computation of the SHAP-score can be reduced in polynomial time to the computation of $\mathsf{H}_{\Pi.}(\cdot,\cdot,\cdot)$.

Suppose then that we wish to compute $\mathsf{SHAP}_{\Pi_p}(C, \mathbf{e}, x)$, for a given deterministic and decomposable circuit C over a set of variables X, probability mapping $p : X \to [0,1]$, entity $\mathbf{e} \in \mathsf{ent}(X)$, and feature $x \in X$. Define

$$\mathsf{Diff}_k(C, \mathbf{e}, x) \ := \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} (\phi_{\Pi_p}(C, \mathbf{e}, S \cup \{x\}) - \phi_{\Pi_p}(C, \mathbf{e}, S)),$$

and let $n = |X|$. We then have, by definition, that

$$\begin{aligned}
\mathsf{SHAP}_{\Pi_p}(C, \mathbf{e}, x) \ &= \sum_{S \subseteq X \setminus \{x\}} \frac{|S|!(n - |S| - 1)!}{n!} (\phi_{\Pi_p}(C, \mathbf{e}, S \cup \{x\}) - \phi_{\Pi_p}(C, \mathbf{e}, S)) \\
&= \sum_{k=0}^{n-1} \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \frac{k!(n-k-1)!}{n!} (\phi_{\Pi_p}(C, \mathbf{e}, S \cup \{x\}) - \phi_{\Pi_p}(C, \mathbf{e}, S)) \\
&= \sum_{k=0}^{n-1} \frac{k!(n-k-1)!}{n!} \mathsf{Diff}_k(C, \mathbf{e}, x). \qquad (3)
\end{aligned}$$

Observe that all arithmetical terms (such as $k!$ or $n!$) can be computed in polynomial time: this is simply because n is given in unary, as it is bounded by the size of the cir-

cuit. Therefore, it is good enough to show how to compute in polynomial time the quantities $\mathsf{Diff}_k(C, \mathbf{e}, x)$ for each $k \in \{0, \ldots, n-1\}$. By definition of $\phi_{\Pi_\cdot}(\cdot,\cdot,\cdot)$ we have that

$$\mathsf{Diff}_k(C, \mathbf{e}, x) = \Big[\sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\})] \Big] - \Big[\sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)] \Big]. \quad (4)$$

In this expression, let L and R be the left- and right-hand side terms in the subtraction. Looking closer at R, we have by standard properties of conditional expectation that

$$\begin{aligned} R &= \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)] \\ &= \mathrm{p}(x) \cdot \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S) \text{ and } \mathbf{e}'(x) = 1] \\ &\qquad + (1 - \mathrm{p}(x)) \cdot \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S) \text{ and } \mathbf{e}'(x) = 0]. \end{aligned}$$

To continue, we need to introduce some notation. For a set of features X and $S \subseteq X$, we write $\mathrm{p}_{|S} : S \to [0,1]$ for the mapping that is the restriction of p to S, and $\Pi_{\mathrm{p}_{|S}} : \mathsf{ent}(S) \to [0,1]$ for the corresponding product distribution on $\mathsf{ent}(S)$. Similarly, for an entity $\mathbf{e} \in \mathsf{ent}(X)$ and $S \subseteq X$, let $\mathbf{e}_{|S}$ be the entity over S that is obtained by restricting $\mathbf{e}$ to the domain S (that is, formally $\mathbf{e}_{|S} \in \mathsf{ent}(S)$ and $\mathbf{e}_{|S}(y) \coloneqq \mathbf{e}(y)$ for every $y \in S$). Now, remembering from Section 2.3 the definition of C_{+x} and C_{-x}, we obtain that

$$\begin{aligned} R &= p(x) \cdot \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}'' \sim \Pi_{p_{|X \setminus \{x\}}}}[C_{+x}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|X \setminus \{x\}}, S)] \\ &\qquad + (1 - p(x)) \cdot \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}'' \sim \Pi_{p_{|X \setminus \{x\}}}}[C_{-x}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|X \setminus \{x\}}, S)] \\ &= p(x) \cdot \mathsf{H}_{\Pi_{p_{|X \setminus \{x\}}}}(C_{+x}, \mathbf{e}_{|X \setminus \{x\}}, k) + (1 - p(x)) \cdot \mathsf{H}_{\Pi_{p_{|X \setminus \{x\}}}}(C_{-x}, \mathbf{e}_{|X \setminus \{x\}}, k), \end{aligned} \quad (5)$$

where the last equality is obtained simply by using the definition of $\mathsf{H}_{\Pi_\cdot}(\cdot,\cdot,\cdot)$. Hence, if we could compute in polynomial time $H_{\Pi_\cdot}(\cdot,\cdot,\cdot)$ for deterministic and decomposable Boolean circuits, then we could compute R in polynomial time as C_{+x} and C_{-x} can be computed in linear time from C, and they are deterministic and decomposable Boolean circuits as well.

We now inspect the term L, which we recall is

$$L = \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\})].$$

Observe that, for $S \subseteq X \setminus \{x\}$ and $\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\})$, it holds that

$$C(\mathbf{e}') \ = \begin{cases} C_{+x}(\mathbf{e}'_{|X\setminus\{x\}}) & \text{if } \mathbf{e}(x) = 1 \\ C_{-x}(\mathbf{e}'_{|X\setminus\{x\}}) & \text{if } \mathbf{e}(x) = 0 \end{cases}.$$

Therefore, if $\mathbf{e}(x) = 1$, we have that

$$\begin{aligned} L \ &= \sum_{\substack{S\subseteq X\setminus\{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}''\sim\Pi_{p_{|X\setminus\{x\}}}}[C_{+x}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|X\setminus\{x\}}, S)] \\ &= \mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(C_{+x}, \mathbf{e}_{|X\setminus\{x\}}, k) \end{aligned} \tag{6}$$

whereas if $\mathbf{e}(x) = 0$, we have that

$$L \ = \ \mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(C_{-x}, \mathbf{e}_{|X\setminus\{x\}}, k). \tag{7}$$

Hence, again, if we were able to compute in polynomial time $\mathsf{H}_{\Pi_\cdot}(\cdot,\cdot,\cdot)$ for deterministic and decomposable Boolean circuits, then we could compute L in polynomial time (as deterministic and decomposable Boolean circuits C_{+x} and C_{-x} can be computed in linear time from C). But then we deduce from (4) that $\mathsf{Diff}_k(C, \mathbf{e}, x)$ could be computed in polynomial time for each $k \in \{0, \ldots, n-1\}$, from which we have that $\mathsf{SHAP}_{\Pi_p}(C, \mathbf{e}, x)$ could be computed in polynomial time (by Equation (3)), therefore concluding the existence of the reduction claimed in this section.

3.1.2 Computing $\mathsf{H}_{\Pi_\cdot}(\cdot,\cdot,\cdot)$ in polynomial time

We now take care of the second part of the proof of tractability, i.e., proving that computing $\mathsf{H}_{\Pi_\cdot}(\cdot,\cdot,\cdot)$ for deterministic and decomposable Boolean circuits can be done in polynomial time. Formally, in this section we prove the following lemma:

Lemma 4 *The following problem can be solved in polynomial time. Given as input a deterministic and decomposable Boolean circuit C over a set of variables X, rational probability values $\mathrm{p}(x)$ for each $x \in X$, an entity $\mathbf{e} \in \mathsf{ent}(X)$, and a natural number $k \leq |X|$, compute the quantity $\mathsf{H}_{\Pi_p}(C, \mathbf{e}, k)$.*

We first perform two preprocessing steps on C.

Rewriting to fan-in 2. First, we modify the circuit C so that the fan-in of every $\vee$- and $\wedge$-gate is exactly 2. This can be done in linear time simply by rewriting every $\wedge$-gate (resp., and $\vee$-gate) of fan-in $m > 2$ with a chain of $m-1$ $\wedge$-gates (resp., $\vee$-gates) of fan-in 2, and by attaching to each $\wedge$ or $\vee$-gate of fan-in 1 a constant gate of the appropriate type. It is clear that the resulting Boolean circuit is deterministic and decomposable. Hence, from now on we assume that the fan-in of every $\vee$- and $\wedge$-gate of C is exactly 2.

Smoothing the circuit. A deterministic and decomposable circuit C is *smooth* (Darwiche, 2001; Shih et al., 2019b) if for every $\vee$-gate g and input gates g_1, g_2 of g, we

have that $\mathsf{var}(g_1) = \mathsf{var}(g_2)$ (we call such an $\vee$-gate smooth). Recall that by the previous paragraph, we assume that the fan-in of every $\vee$-gate is exactly 2. We then repeat the following operation until C becomes smooth. For an $\vee$-gate g of C having two input gates g_1, g_2 violating the smoothness condition, define $S_1 := \mathsf{var}(g_1) \setminus \mathsf{var}(g_2)$ and $S_2 := \mathsf{var}(g_2) \setminus \mathsf{var}(g_1)$, and let d_{S_1}, d_{S_2} be Boolean circuits defined as follows. If $S_1 = \emptyset$, then d_{S_1} consist of the single constant 1-gate. Otherwise, d_{S_1} encodes the propositional formula $\wedge_{x \in S_1}(x \vee \neg x)$ but it is constructed as a circuit in such a way that every $\wedge$- and $\vee$-gate has fan-in exactly 2. Boolean circuit d_{S_2} is constructed exactly as d_{S_1} but considering the set of variables S_2 instead of S_1. Observe that $\mathsf{var}(d_{S_1}) = S_1$, $\mathsf{var}(d_{S_2}) = S_2$, that d_{S_1} and d_{S_2} always evaluate to 1, and that all $\vee$-gates appearing in d_{S_1} and in d_{S_2} are deterministic. Then, we transform g into a smooth $\vee$-gate by replacing gate g_1 by a decomposable $\wedge$-gate $(g_1 \wedge d_{S_2})$, and gate g_2 by a decomposable $\wedge$-gate $(g_2 \wedge d_{S_1})$. Clearly, this does not change the Boolean classifier computed, and g is again deterministic because g_1 and $(g_1 \wedge d_{S_2})$ (resp., g_2 and $(g_2 \wedge d_{S_1})$) capture the same Boolean classifier. Moreover, since $\mathsf{var}(g_1 \wedge d_{S_2}) = \mathsf{var}(g_2 \wedge d_{S_1}) = \mathsf{var}(g_1) \cup \mathsf{var}(g_2)$, we have that g is now smooth. Therefore, the circuit that we obtain after this operation is equivalent to the one we started from, is again deterministic and decomposable, and contains one less non-smooth $\vee$-gate. Hence, by repeating this operation for each non-smooth $\vee$-gate, which we can do in polynomial time, we obtain an equivalent smooth deterministic and decomposable circuit where each $\vee$- and $\wedge$-gate has fan-in exactly 2. Thus, from now on we also assume that C is smooth.

For a gate g of C, let R_g be the Boolean circuit over $\mathsf{var}(g)$ that is defined by considering the subgraph of C induced by the set of gates g' in C for which there exists a path from g' to g in C.[4] Notice that R_g is a deterministic and decomposable Boolean circuit with output gate g. Moreover, for a gate g and natural number $\ell \leq |\mathsf{var}(g)|$, define $\alpha_g^\ell := \mathsf{H}_{\Pi_{p_{|\mathsf{var}(g)}}}(R_g, \mathbf{e}_{|\mathsf{var}(g)}, \ell)$, which we recall is equal, by definition of $\mathsf{H}_{\Pi_\cdot}(\cdot,\cdot,\cdot)$, to

$$\mathsf{H}_{\Pi_{p_{|\mathsf{var}(g)}}}(R_g, \mathbf{e}_{|\mathsf{var}(g)}, \ell) \;=\; \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S| = \ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}}[R_g(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)].$$

We will show how to compute all the values α_g^ℓ for every gate g of C and $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ in polynomial time. This will conclude the proof of Lemma 4 since, for the output gate g_{out} of C, we have that $\alpha_{g_{\mathsf{out}}}^k = \mathsf{H}_{\Pi_p}(C, \mathbf{e}, k)$. Next we explain how to compute these values by bottom-up induction on C.

Variable gate. g is a variable gate with label $y \in X$, so that $\mathsf{var}(g) = \{y\}$. Then for $\mathbf{e}' \in \mathsf{ent}(\{y\})$ we have $R_g(\mathbf{e}') = \mathbf{e}'(y)$, therefore

$$\begin{aligned} \alpha_g^0 &= \sum_{\substack{S \subseteq \{y\} \\ |S| = 0}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\{y\}}}}[\mathbf{e}'(y) \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\{y\}}, S)] \\ &= \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\{y\}}}}[\mathbf{e}'(y) \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\{y\}}, \emptyset)] \end{aligned}$$

4. The only difference between R_g and C_g (defined in Section 2) is that we formally regard R_g as a Boolean classifier over $\mathsf{var}(g)$, while we formally regarded C_g as a Boolean classifier over X.

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\{y\}}}}[\mathbf{e}'(y)] \\
&= 1 \cdot p(y) + 0 \cdot (1 - p(y)) \\
&= p(y) \qquad (8)
\end{aligned}
$$

and

$$
\begin{aligned}
\alpha_g^1 &= \sum_{\substack{S \subseteq \{y\} \\ |S|=1}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\{y\}}}}[\mathbf{e}'(y) \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\{y\}}, S)] \\
&= \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\{y\}}}}[\mathbf{e}'(y) \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\{y\}}, \{y\})] \\
&= \mathbf{e}(y). \qquad (9)
\end{aligned}
$$

Constant gate. g is a constant gate with label $a \in \{0, 1\}$, and $\mathsf{var}(g) = \emptyset$. We recall the mathematical convention that there is a unique function with the empty domain and, hence, a unique entity over $\emptyset$. But then

$$
\begin{aligned}
\alpha_g^0 &= \sum_{\substack{S \subseteq \emptyset \\ |S|=0}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\emptyset}}}[a \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\emptyset}, S)] \\
&= \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\emptyset}}}[a \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\emptyset}, \emptyset)] \\
&= a. \qquad (10)
\end{aligned}
$$

¬-gate. g is a ¬-gate with input gate g'. Notice that $\mathsf{var}(g) = \mathsf{var}(g')$. Then, since for $\mathbf{e}' \in \mathsf{ent}(\mathsf{var}(g))$ we have that $R_g(\mathbf{e}') = 1 - R_{g'}(\mathbf{e}')$, we have

$$
\alpha_g^\ell = \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S|=\ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}}[1 - R_{g'}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)].
$$

By linearity of expectations we deduce that

$$
\begin{aligned}
\alpha_g^\ell &= \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S|=\ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}}[1 \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \\
&\qquad - \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S|=\ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}}[R_{g'}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \\
&= \Big(\sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S|=\ell}} 1 \Big) - \alpha_{g'}^\ell \\
&= \binom{|\mathsf{var}(g)|}{\ell} - \alpha_{g'}^\ell \qquad (11)
\end{aligned}
$$

for every $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$. By induction, the values $\alpha_{g'}^\ell$ for $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ have already been computed. Thus, we can compute all the values α_g^ℓ for $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ in polynomial time.

∨-gate. g is an $\vee$-gate. By assumption, recall that g is deterministic, smooth, and has fan-in exactly 2. Let g_1 and g_2 be the input gates of g, and recall that $\mathsf{var}(g_1) = \mathsf{var}(g_2) = \mathsf{var}(g)$, because g is smooth. Given that g is deterministic, observe that for every $\mathbf{e}' \in \mathsf{ent}(\mathsf{var}(g))$ we have $R_g(\mathbf{e}') = R_{g_1}(\mathbf{e}') + R_{g_2}(\mathbf{e}')$. But then for $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ we have

$$\begin{aligned}
\alpha_g^\ell &= \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S| = \ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}} [R_{g_1}(\mathbf{e}') + R_{g_2}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \\
&= \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S| = \ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}} [R_{g_1}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \\
&\qquad + \sum_{\substack{S \subseteq \mathsf{var}(g) \\ |S| = \ell}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}} [R_{g_2}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \\
&= \alpha_{g_1}^\ell + \alpha_{g_2}^\ell, \qquad (12)
\end{aligned}$$

where the second equality is by linearity of the expectation, and the last equality is valid because g is smooth (in particular, we have that $\mathsf{var}(g_1) = \mathsf{var}(g_2) = \mathsf{var}(g)$). By induction, the values $\alpha_{g_1}^\ell$ and $\alpha_{g_2}^\ell$, for each $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$, have already been computed. Therefore, we can compute all the values α_g^ℓ for $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ in polynomial time.

∧-gate. g is an $\wedge$-gate. By assumption, recall that g is decomposable and has fan-in exactly 2. Let g_1 and g_2 be the input gates of g. For $\mathbf{e}' \in \mathsf{ent}(\mathsf{var}(g))$ we have that $R_g(\mathbf{e}') = R_{g_1}(\mathbf{e}'_{|\mathsf{var}(g_1)}) \cdot R_{g_2}(\mathbf{e}'_{|\mathsf{var}(g_2)})$. Moreover, since $\mathsf{var}(g) = \mathsf{var}(g_1) \cup \mathsf{var}(g_2)$ and $\mathsf{var}(g_1) \cap \mathsf{var}(g_2) = \emptyset$ (because g is decomposable), observe that every $S \subseteq \mathsf{var}(g)$ can be uniquely decomposed into $S_1 \subseteq \mathsf{var}(g_1)$, $S_2 \subseteq \mathsf{var}(g_2)$ such that $S = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$. Therefore, for $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ we have

$$\begin{aligned}
\alpha_g^\ell = \sum_{\substack{S_1 \subseteq \mathsf{var}(g_1) \\ |S_1| \leq \ell}} \sum_{\substack{S_2 \subseteq \mathsf{var}(g_2) \\ |S_2| = \ell - |S_1|}} \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}} [R_{g_1}(\mathbf{e}'_{|\mathsf{var}(g_1)}) \cdot R_{g_2}(\mathbf{e}'_{|\mathsf{var}(g_2)}) \mid \\
\mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S_1 \cup S_2)].
\end{aligned}$$

But, by definition of the product distribution $\Pi_{p_{|\mathsf{var}(g)}}$ and because g is decomposable, we have that $R_{g_1}(\mathbf{e}'_{|\mathsf{var}(g_1)})$ and $R_{g_2}(\mathbf{e}'_{|\mathsf{var}(g_2)})$ are independent random variables, hence we deduce

$$\begin{aligned}
\alpha_g^\ell &= \sum_{\substack{S_1 \subseteq \mathsf{var}(g_1) \\ |S_1| \leq \ell}} \sum_{\substack{S_2 \subseteq \mathsf{var}(g_2) \\ |S_2| = \ell - |S_1|}} \Big(\mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}} [R_{g_1}(\mathbf{e}'_{|\mathsf{var}(g_1)}) \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S_1 \cup S_2)] \\
&\qquad \cdot \mathbb{E}_{\mathbf{e}' \sim \Pi_{p_{|\mathsf{var}(g)}}} [R_{g_2}(\mathbf{e}'_{|\mathsf{var}(g_2)}) \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S_1 \cup S_2)] \Big). \\
&= \sum_{\substack{S_1 \subseteq \mathsf{var}(g_1) \\ |S_1| \leq \ell}} \sum_{\substack{S_2 \subseteq \mathsf{var}(g_2) \\ |S_2| = \ell - |S_1|}} \Big(\mathbb{E}_{\mathbf{e}'' \sim \Pi_{p_{|\mathsf{var}(g_1)}}} [R_{g_1}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g_1)}, S_1)]
\end{aligned}$$

$$\cdot\, \mathbb{E}_{\mathbf{e}''\sim\Pi_{p_{|\mathsf{var}(g_2)}}}[R_{g_2}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g_2)}, S_2)]\Big),$$

where the last equality is simply by definition of the product distributions, and because $R_{g_1}(\mathbf{e}'_{|\mathsf{var}(g_1)})$ is independent of the value $\mathbf{e}'_{|\mathsf{var}(g_2)}$, and similarly for $R_{g_2}(\mathbf{e}'_{|\mathsf{var}(g_2)})$. But then, using the convention that $\alpha_{g_i}^{\ell_i} = 0$ when $\ell_i > |\mathsf{var}(g_i)|$, for $i = 1, 2$, we obtain that

$$\begin{aligned}
\alpha_g^\ell &= \sum_{\substack{S_1\subseteq\mathsf{var}(g_1)\\ |S_1|\le\ell}} \mathbb{E}_{\mathbf{e}''\sim\Pi_{p_{|\mathsf{var}(g_1)}}}[R_{g_1}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g_1)}, S_1)] \cdot \\
&\qquad\qquad \sum_{\substack{S_2\subseteq\mathsf{var}(g_2)\\ |S_2|=\ell-|S_1|}} \mathbb{E}_{\mathbf{e}''\sim\Pi_{p_{|\mathsf{var}(g_2)}}}[R_{g_2}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g_2)}, S_2)] \\
&= \sum_{\substack{S_1\subseteq\mathsf{var}(g_1)\\ |S_1|\le\ell}} \mathbb{E}_{\mathbf{e}''\sim\Pi_{p_{|\mathsf{var}(g_1)}}}[R_{g_1}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g_1)}, S_1)] \cdot \alpha_{g_2}^{\ell-|S_1|} \\
&= \sum_{\ell_1=0}^{\ell} \alpha_{g_2}^{\ell-\ell_1} \cdot \sum_{\substack{S_1\subseteq\mathsf{var}(g_1)\\ |S_1|=\ell_1}} \mathbb{E}_{\mathbf{e}''\sim\Pi_{p_{|\mathsf{var}(g_1)}}}[R_{g_1}(\mathbf{e}'') \mid \mathbf{e}'' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g_1)}, S_1)] \\
&= \sum_{\ell_1=0}^{\ell} \alpha_{g_2}^{\ell-\ell_1} \cdot \alpha_{g_1}^{\ell_1} \\
&= \sum_{\substack{\ell_1\in\{0,\ldots,\min(\ell,|\mathsf{var}(g_1)|)\}\\ \ell_2\in\{0,\ldots,\min(\ell,|\mathsf{var}(g_2)|)\}\\ \ell_1+\ell_2=\ell}} \alpha_{g_1}^{\ell_1} \cdot \alpha_{g_2}^{\ell_2}. \qquad (13)
\end{aligned}$$

By induction, the values $\alpha_{g_1}^{\ell_1}$ and $\alpha_{g_2}^{\ell_2}$, for each $\ell_1 \in \{0, \ldots, |\mathsf{var}(g_1)|\}$ and $\ell_2 \in \{0, \ldots, |\mathsf{var}(g_2)|\}$, have already been computed. Therefore, we can compute all the values α_g^ℓ for $\ell \in \{0, \ldots, |\mathsf{var}(g)|\}$ in polynomial time.

This concludes the proof of Lemma 4 and, hence, the proof that SHAP-score can be computed in polynomial time for our circuits.

3.1.3 Extracting an algorithm from the proof

From the previous proof, it is possible to extract Algorithm 1, which is more amenable to implementation. The main idea of this algorithm is that values γ_g^ℓ correspond to values α_ℓ^g of the proof for the circuit D_{+x}, while values δ_g^ℓ correspond to values α_g^ℓ of the proof for the circuit D_{-x}. In lines 3–27, these values are computed by bottom-up induction over the circuit D, following the relations that we obtained in Equations (8)–(13) (but specialized to the circuits D_{+x} and D_{-x}, as can be seen from lines 6–8, and by the fact that indices are always in $\{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$). Hence, it only remains to show that the returned value of the algorithm is correct. To see that, observe that we can rewrite Equations (6) and (7) into

$$L = \mathbf{e}(x) \cdot \mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(D_{+x}, \mathbf{e}_{|X\setminus\{x\}}, k) + (1 - \mathbf{e}(x)) \cdot \mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(D_{-x}, \mathbf{e}_{|X\setminus\{x\}}, k),$$

Algorithm 1: SHAP-scores for deterministic and decomposable Boolean circuits (intermediate)

Input : Deterministic and decomposable Boolean circuit C over features X with output gate g_{out}, rational probability values $\mathrm{p}(x)$ for all $x \in X$, entity $\mathbf{e} \in \mathsf{ent}(X)$, and feature $x \in X$.

Output: The value $\mathsf{SHAP}(C, \mathbf{e}, x)$ under the probability distribution Π_{p}.

Transform C into an equivalent smooth circuit D where each $\vee$-gate and $\wedge$-gate has fan-in exactly 2;

Compute the set $\mathsf{var}(g)$ for every gate g in D;

Compute values γ_g^ℓ and δ_g^ℓ for every gate g in D and $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ by bottom-up induction on D as follows:

- **if** *g is a constant gate with label $a \in \{0,1\}$* **then**
  - $\gamma_g^0, \delta_g^0 \leftarrow a$;
- **else if** *g is a variable gate with $\mathsf{var}(g) = \{x\}$* **then**
  - $\gamma_g^0 \leftarrow 1$;
  - $\delta_g^0 \leftarrow 0$;
- **else if** *g is a variable gate with $\mathsf{var}(g) = \{y\}$ and $y \neq x$* **then**
  - $\gamma_g^0, \delta_g^0 \leftarrow \mathrm{p}(y)$;
  - $\gamma_g^1, \delta_g^1 \leftarrow \mathbf{e}(y)$;
- **else if** *g is a $\neg$-gate with input gate g'* **then**
  - **for** $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ **do**
    - $\gamma_g^\ell \leftarrow \binom{|\mathsf{var}(g)\setminus\{x\}|}{\ell} - \gamma_{g'}^\ell$;
    - $\delta_g^\ell \leftarrow \binom{|\mathsf{var}(g)\setminus\{x\}|}{\ell} - \delta_{g'}^\ell$;
  - **end**
- **else if** *g is an $\vee$-gate with input gates g_1, g_2* **then**
  - **for** $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ **do**
    - $\gamma_g^\ell \leftarrow \gamma_{g_1}^\ell + \gamma_{g_2}^\ell$;
    - $\delta_g^\ell \leftarrow \delta_{g_1}^\ell + \delta_{g_2}^\ell$;
  - **end**
- **else if** *g is an $\wedge$-gate with input gates g_1, g_2* **then**
  - **for** $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ **do**
    - $\gamma_g^\ell \leftarrow \sum_{\substack{\ell_1 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)\setminus\{x\}|)\} \\ \ell_2 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)\setminus\{x\}|)\} \\ \ell_1 + \ell_2 = \ell}} \gamma_{g_1}^{\ell_1} \cdot \gamma_{g_2}^{\ell_2}$;
    - $\delta_g^\ell \leftarrow \sum_{\substack{\ell_1 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)\setminus\{x\}|)\} \\ \ell_2 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)\setminus\{x\}|)\} \\ \ell_1 + \ell_2 = \ell}} \delta_{g_1}^{\ell_1} \cdot \delta_{g_2}^{\ell_2}$;
  - **end**

end

return $\displaystyle\sum_{k=0}^{|X|-1} \frac{k!\,(|X|-k-1)!}{|X|!} \cdot \left[(\mathbf{e}(x) - \mathrm{p}(x))(\gamma_{g_{\mathsf{out}}}^k - \delta_{g_{\mathsf{out}}}^k)\right]$;

which works no matter the value of $\mathbf{e}(x) \in \{0,1\}$. We can then directly combine this expression for L with Equation (4) and the expression of R from Equation (5) to obtain

$$\begin{aligned}\mathsf{Diff}_k(D,\mathbf{e},x) &= L - R \\ &= (\mathbf{e}(x) - \mathrm{p}(x))\big[\mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(D_{+x},\mathbf{e}_{|X\setminus\{x\}},k) - \mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(D_{-x},\mathbf{e}_{|X\setminus\{x\}},k)\big].\end{aligned}$$

But the rightmost factor is precisely $(\gamma^k_{g_{\mathsf{out}}} - \delta^k_{g_{\mathsf{out}}})$ in Algorithm 1, so that the returned value is indeed correct by Equation (3).

Example 2 We describe in this example a complete execution of Algorithm 1 over the deterministic and decomposable Boolean circuit C in depicted Figure 1 (see Example 1). More precisely, we show in Figure 2 how $\mathsf{SHAP}(C,\mathbf{e},\mathbf{nf})$ is computed under the uniform distribution (that is, $\mathrm{p}(\mathbf{fg}) = \mathrm{p}(\mathbf{dtr}) = \mathrm{p}(\mathbf{nf}) = \mathrm{p}(\mathbf{na}) = 1/2$), and assuming that $\mathbf{e}(\mathbf{fg}) = 1$, $\mathbf{e}(\mathbf{dtr}) = 0$, $\mathbf{e}(\mathbf{nf}) = 1$, and $\mathbf{e}(\mathbf{na}) = 1$. Notice that $C(\mathbf{e}) = 1$.

In the first step of the algorithm, the gate in gray in Figure 2 is added to C so that the fan-in of every $\vee$- and $\wedge$-gate is exactly 2, and the gates in blue in Figure 2 are added to obtain a deterministic, decomposable and smooth circuit D (in particular, for every $\vee$-gate g of D with input gates g_1 and g_2, it holds that $\mathsf{var}(g_1) = \mathsf{var}(g_2)$). Then in the main loop of the algorithm, the values of γ^i_g and δ^i_g are computed in a bottom-up fashion for every gate g and $i \in \{0,\ldots,|\mathsf{var}(g) \setminus \{\mathbf{nf}\}|\}$. Finally, assuming that g_{out} is the top $\wedge$-gate in D, the value $\mathsf{SHAP}(C,\mathbf{e},\mathbf{nf})$ is computed as follows:

$$\begin{aligned}\mathsf{SHAP}(C,\mathbf{e},\mathbf{nf}) &= \sum_{k=0}^{3} \frac{k!\,(3-k)!}{4!} \cdot \big[(\mathbf{e}(\mathbf{nf}) - \mathrm{p}(\mathbf{nf}))(\gamma^k_{g_{\mathsf{out}}} - \delta^k_{g_{\mathsf{out}}})\big] \\ &= \frac{1}{2}\cdot\left(\frac{1}{4}\cdot\frac{1}{8} + \frac{1}{12}\cdot\frac{3}{4} + \frac{1}{12}\cdot\frac{3}{2} + \frac{1}{4}\cdot 1\right) \\ &= \frac{15}{64}.\end{aligned}$$

As a final comment, we notice that by following the same procedure it can be shown that $\mathsf{SHAP}(C,\mathbf{e},\mathbf{fg}) = 23/64$, $\mathsf{SHAP}(C,\mathbf{e},\mathbf{dtr}) = -9/64$, and $\mathsf{SHAP}(C,\mathbf{e},\mathbf{na}) = 15/64$. □

3.2 An optimized version of the algorithm

In this section, we present an optimized version of the algorithm to compute the SHAP-score. The observation behind this algorithm is that it is possible to bypass the smoothing step and directly work on the input circuit C.[5] Our procedure can be found in Algorithm 2.

Observe that Algorithm 2 is identical to Algorithm 1, except that: (1) we do not smooth the circuit on line 1; and (2) the expressions for $\vee$-gates on lines 19 and 20 have changed. Before showing the correctness of Algorithm 2, notice that the smoothing step on line 1 of Algorithm 1 takes time $O(|C|\cdot|X|)$, as described in Section 3.1.2, and the resulting circuit, called D, has size $O(|C|\cdot|X|)$. As Algorithm 2 does not execute such a smoothing step, it

5. We first presented a version that used smoothing because it made the tractability proof easier to understand.

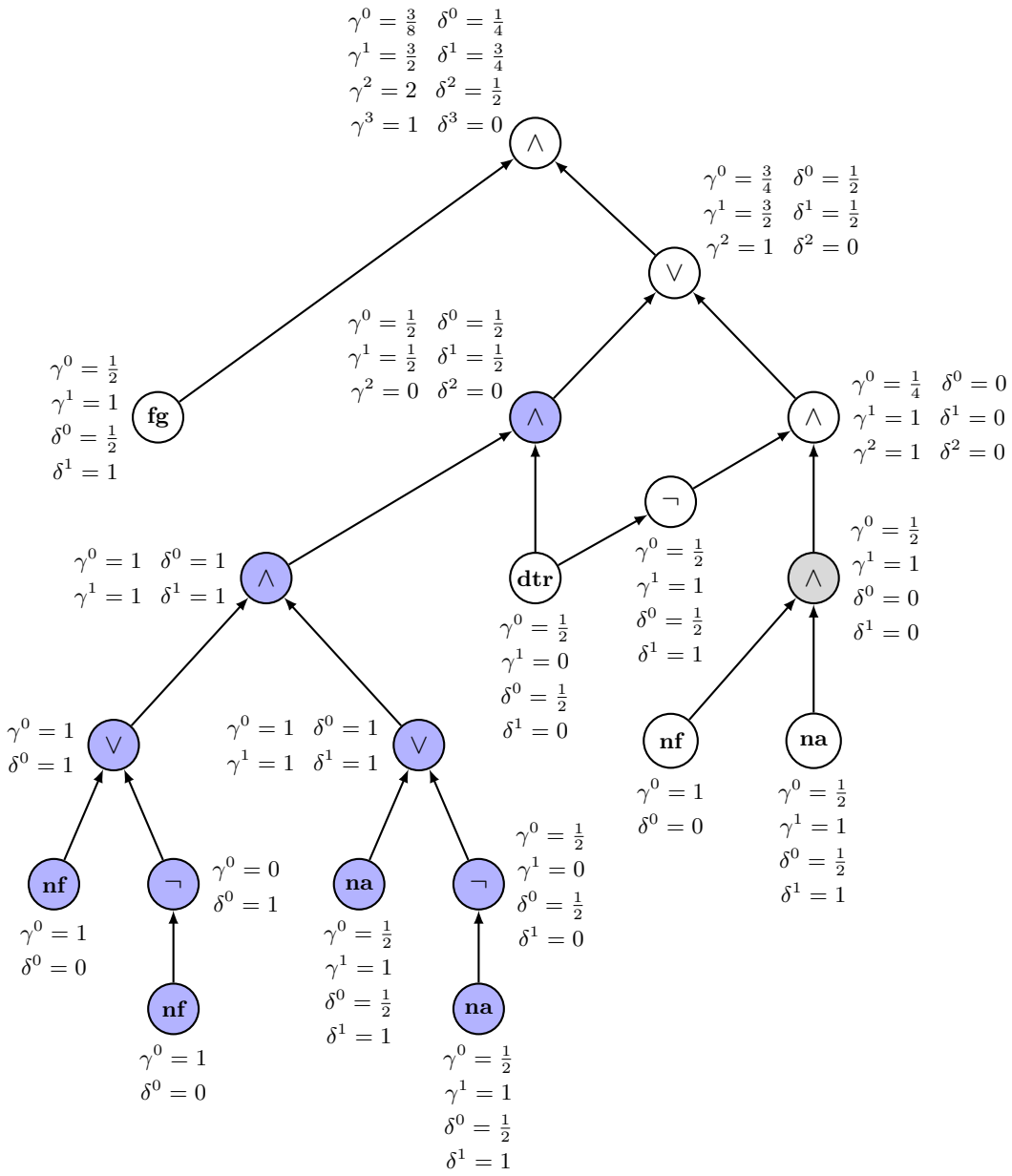


Figure 2: Execution of Algorithm 1 over the deterministic and decomposable Boolean circuit depicted in Figure 1.

works directly with a circuit of size $O(|C|)$, obtained after preprocessing input circuit C to ensure that all $\vee$- and $\wedge$-gate have fan-in 2, and it has a lower complexity.

The only thing that changed between Algorithm 1 and 2 is how we treat an $\vee$-gate, which can now be non-smoothed. Therefore, to show that Algorithm 2 is correct, we need to revisit Equation (12). The relation

$$\begin{aligned}\alpha_g^\ell &= \sum_{\substack{S\subseteq \mathsf{var}(g)\\ |S|=\ell}} \mathbb{E}_{\mathbf{e}'\sim\Pi_{p_{|\mathsf{var}(g)}}}[R_{g_1}(\mathbf{e}') + R_{g_2}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \\ &= \sum_{\substack{S\subseteq \mathsf{var}(g)\\ |S|=\ell}} \mathbb{E}_{\mathbf{e}'\sim\Pi_{p_{|\mathsf{var}(g)}}}[R_{g_1}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)] \qquad (14)\\ &\qquad + \sum_{\substack{S\subseteq \mathsf{var}(g)\\ |S|=\ell}} \mathbb{E}_{\mathbf{e}'\sim\Pi_{p_{|\mathsf{var}(g)}}}[R_{g_2}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|\mathsf{var}(g)}, S)]\end{aligned}$$

is still true for an arbitrary deterministic $\vee$-gate g with children g_1, g_2, however this is not equal to $\alpha_{g_1}^\ell + \alpha_{g_2}^\ell$ anymore when g is not smooth. This is because, in this case, one of $\mathsf{var}(g_1)$ or $\mathsf{var}(g_2)$ (or both) is not equal to $\mathsf{var}(g)$. Assuming without loss of generality that we have $\mathsf{var}(g_1) \neq \mathsf{var}(g)$, by induction hypothesis we have $\alpha_{g_1}^\ell = \mathsf{H}_{\Pi_{p_{|\mathsf{var}(g_1)}}}(R_{g_1}, \mathbf{e}_{|\mathsf{var}(g_1)}, \ell)$, which is not the expression (14) that we obtain above. To fix this, we will do as if the gate had been smoothed, by considering the virtual circuits d_{S_1} and d_{S_2} that we use in the smoothing process, but without materializing them. We recall that $S_1 \coloneqq \mathsf{var}(g_1) \setminus \mathsf{var}(g_2)$ and $S_2 \coloneqq \mathsf{var}(g_2) \setminus \mathsf{var}(g_1)$, and that d_{S_1} and d_{S_2} are Boolean circuits that always evaluate to 1 over sets of variables S_1 and S_2, respectively. Let g_1' and g_2' be the output gate of those circuits. We will show how to compute the values $\alpha_{g_1'}^{\ell_1}$ and $\alpha_{g_2'}^{\ell_2}$, for $\ell_1 \in \{0, \ldots, |S_1|\}$ and $\ell_2 \in \{0, \ldots, |S_2|\}$, directly with a closed form expression and use Equation (13) for $\wedge$-gates to fix the algorithm. We have

$$\alpha_{g_1'}^{\ell_1} = \mathsf{H}_{\Pi_{p_{|\mathsf{var}(g_1')}}}(R_{g_1'}, \mathbf{e}_{|\mathsf{var}(g_1')}, \ell) = \sum_{\substack{S\subseteq S_1\\ |S|=\ell_1}} \mathbb{E}_{\mathbf{e}'\sim\Pi_{p_{|S_1}}}[R_{g_1'}(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|S_1}, S)] =$$

$$\sum_{\substack{S\subseteq S_1\\ |S|=\ell_1}} \mathbb{E}_{\mathbf{e}'\sim\Pi_{p_{|S_1}}}[1 \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|S_1}, S)] = \sum_{\substack{S\subseteq S_1\\ |S|=\ell_1}} 1 = \binom{|S_1|}{\ell_1}.$$

Hence, by virtually considering that the circuit has been smoothed (that is, that we replaced gate g_1 with $(g_1 \wedge d_{S_2})$ and gate g_2 with $(g_2 \wedge d_{S_1})$), and by using the relation for $\wedge$-gates we can correct Equation (12) as follows:

$$\begin{aligned}\alpha_g^\ell = &\sum_{\ell' \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)|)\}} \alpha_{g_1}^{\ell'} \cdot \binom{|\mathsf{var}(g_2) \setminus \mathsf{var}(g_1)|}{\ell - \ell'} \\ &+ \sum_{\ell' \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)|)\}} \alpha_{g_2}^{\ell'} \cdot \binom{|\mathsf{var}(g_1) \setminus \mathsf{var}(g_2)|}{\ell - \ell'}.\end{aligned}$$

And this is precisely what we use in Algorithm 2 for $\vee$-gates, so this concludes the proof.

We note that, if we ignore the complexity of arithmetic operations (that is, if we consider that arithmetic operations over rationals take constant time and that rationals can be stored in constant space), Algorithm 2 runs in time $O(|C| \cdot |X|^2)$.

4. Extension to Non-Binary Deterministic and Decomposable Circuits

In this section, we show how to extend the result of the previous section to non-binary classifiers. First, we need to redefine the notions of entities, product distributions and SHAP-score to account for non-binary features. We point out that these new definitions are to be considered for this section only, as in the rest of the paper we will again consider binary classifiers. Let X be a finite set of features, and dom be a function that associates to every feature $x \in X$ a finite domain $\mathsf{dom}(x)$. An entity over (X, dom) is a function $\mathbf{e}$ that associates to every feature an element $\mathbf{e}(x) \in \mathsf{dom}(X)$. We denote by $\mathsf{ent}(X, \mathsf{dom})$ the set of all entities over (X, dom). We then consider product distributions on $\mathsf{ent}(X, \mathsf{dom})$ as follows. For every $x \in X$, let $\mathrm{p}_x : \mathsf{dom}(X) \to [0, 1]$ be such that $\sum_{v \in \mathsf{dom}(X)} \mathrm{p}_x(v) = 1$. Then the *product distribution generated by* $(\mathrm{p}_x)_{x \in X}$ is the probability distribution Π_p over $\mathsf{ent}(X, \mathsf{dom})$ such that, for every $\mathbf{e} \in \mathsf{ent}(X, \mathsf{dom})$ we have

$$\Pi_\mathrm{p}(\mathbf{e}) \quad := \quad \prod_{x \in X} \mathrm{p}_x(\mathbf{e}(x)).$$

That is, again, Π_p is the product distribution that is determined by pre-specified marginal distributions, and that makes the features take values independently from each other. A *Boolean classifier* M over X is a function $M : \mathsf{ent}(X, \mathsf{dom}) \to \{0, 1\}$ that maps every entity over X to 0 or 1. The SHAP score over such classifiers is then defined just like in Section 2.2.

Non-binary Boolean circuits. We now define a variant of deterministic and decomposable circuits for non-binary variables, that we will call *deterministic and decomposable non-binary Boolean circuits*. Just like a Boolean circuit captures a set of binary entities (those that satisfy the circuit), a *non-binary Boolean circuit* will capture a set of entities over (X, dom). We define a non-binary Boolean circuit over (X, dom) like a Boolean circuit (recall the definition from Section 2.3), except that each variable gate is now labeled with an equality of the form $x = v$, where $x \in X$ and $v \in \mathsf{dom}(X)$. Such a circuit then maps every entity over X to 0 or 1 in the expected way, and can thus be seen as a Boolean classifier. Again, an $\wedge$-gate is decomposable if for every pair g_1, g_2 of distinct input gates of g, we have that $\mathsf{var}(g_1) \cap \mathsf{var}(g_2) = \emptyset$; an $\vee$-gate is deterministic if for every pair g_1, g_2 of distinct input gates of g there is no entity that satisfies them both; and the circuit is called deterministic and decomposable if all its $\wedge$-gates are decomposable and all its $\vee$-gates are deterministic.

The main result of this section is that we can generalize Theorem 2 to these kind of circuits (we presented the result for Boolean circuits over binary variables first for clarity of presentation):

Theorem 5 *Given as input a set of features X, finite domains $\mathsf{dom}(x)$ for every $x \in X$, a deterministic and decomposable non-binary Boolean circuit C over (X, dom), rational probability values $\mathrm{p}_x(v)$ for every $x \in X$ and $v \in \mathsf{dom}(x)$, an entity $\mathbf{e} \in \mathsf{ent}(X, \mathsf{dom})$, and a feature $x \in X$, the value $\mathsf{SHAP}_{\Pi_\mathrm{p}}(C, \mathbf{e}, x)$ can be computed in polynomial time.*

Algorithm 2: SHAP-scores for deterministic and decomposable Boolean circuits

Input : Deterministic and decomposable Boolean circuit C over features X with output gate g_{out}, rational probability values $\mathrm{p}(x)$ for all $x \in X$, entity $\mathbf{e} \in \mathsf{ent}(X)$, and feature $x \in X$.

Output: The value $\mathsf{SHAP}(C, \mathbf{e}, x)$ under the probability distribution Π_{p}.

Preprocess C so that each $\vee$-gate and $\wedge$-gate has fan-in exactly 2;

Compute the set $\mathsf{var}(g)$ for every gate g in C;

Compute values γ_g^ℓ and δ_g^ℓ for every gate g in C and $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ by bottom-up induction on C as follows:

- **if** *g is a constant gate with label $a \in \{0,1\}$* **then**
  - $\gamma_g^0, \delta_g^0 \leftarrow a$;
- **else if** *g is a variable gate with $\mathsf{var}(g) = \{x\}$* **then**
  - $\gamma_g^0 \leftarrow 1$;
  - $\delta_g^0 \leftarrow 0$;
- **else if** *g is a variable gate with $\mathsf{var}(g) = \{y\}$ and $y \neq x$* **then**
  - $\gamma_g^0, \delta_g^0 \leftarrow \mathrm{p}(y)$;
  - $\gamma_g^1, \delta_g^1 \leftarrow \mathbf{e}(y)$;
- **else if** *g is a $\neg$-gate with input gate g'* **then**
  - **for** $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ **do**
    - $\gamma_g^\ell \leftarrow \binom{|\mathsf{var}(g)\setminus\{x\}|}{\ell} - \gamma_{g'}^\ell$;
    - $\delta_g^\ell \leftarrow \binom{|\mathsf{var}(g)\setminus\{x\}|}{\ell} - \delta_{g'}^\ell$;
  - **end**
- **else if** *g is an $\vee$-gate with input gates g_1, g_2* **then**
  - **for** $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ **do**
    - $\gamma_g^\ell \leftarrow \sum_{\ell' \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)\setminus\{x\}|)\}} \gamma_{g_1}^{\ell'} \cdot \binom{|\mathsf{var}(g_2)\setminus(\mathsf{var}(g_1)\cup\{x\})|}{\ell-\ell'} + \sum_{\ell' \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)\setminus\{x\}|)\}} \gamma_{g_2}^{\ell'} \cdot \binom{|\mathsf{var}(g_1)\setminus(\mathsf{var}(g_2)\cup\{x\})|}{\ell-\ell'}$;
    - $\delta_g^\ell \leftarrow \sum_{\ell' \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)\setminus\{x\}|)\}} \delta_{g_1}^{\ell'} \cdot \binom{|\mathsf{var}(g_2)\setminus(\mathsf{var}(g_1)\cup\{x\})|}{\ell-\ell'} + \sum_{\ell' \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)\setminus\{x\}|)\}} \delta_{g_2}^{\ell'} \cdot \binom{|\mathsf{var}(g_1)\setminus(\mathsf{var}(g_2)\cup\{x\})|}{\ell-\ell'}$;
  - **end**
- **else if** *g is an $\wedge$-gate with input gates g_1, g_2* **then**
  - **for** $\ell \in \{0, \ldots, |\mathsf{var}(g) \setminus \{x\}|\}$ **do**
    - $\gamma_g^\ell \leftarrow \sum_{\substack{\ell_1 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)\setminus\{x\}|)\} \\ \ell_2 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)\setminus\{x\}|)\} \\ \ell_1+\ell_2=\ell}} \gamma_{g_1}^{\ell_1} \cdot \gamma_{g_2}^{\ell_2}$;
    - $\delta_g^\ell \leftarrow \sum_{\substack{\ell_1 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_1)\setminus\{x\}|)\} \\ \ell_2 \in \{0,\ldots,\min(\ell, |\mathsf{var}(g_2)\setminus\{x\}|)\} \\ \ell_1+\ell_2=\ell}} \delta_{g_1}^{\ell_1} \cdot \delta_{g_2}^{\ell_2}$;
  - **end**
- **end**

return $\displaystyle\sum_{k=0}^{|X|-1} \frac{k!\,(|X|-k-1)!}{|X|!} \cdot \big[(\mathbf{e}(x) - \mathrm{p}(x))(\gamma_{g_{\mathsf{out}}}^k - \delta_{g_{\mathsf{out}}}^k)\big]$;

Proof First, notice that, since probability values $\mathrm{p}_x(v)$ for every $x \in X$ and $v \in \mathsf{dom}(x)$ are anyway given as part of the input, it can safely be considered that $\mathsf{dom}(x)$ is of linear size. We then go through the proof of Theorem 2 and only explain what changes. For $x \in X$ and $v \in \mathsf{dom}(x)$, we denote by $C_{x=v}$ the non-binary Boolean circuit that is obtained from C by replacing every variable gate that is labeled with $x = v$ by a constant 1-gate, and every variable gate that is labeled with $x = v'$ for $v' \neq v$ by a constant 0-gate (it is clear that $C_{x=v}$ is again deterministic and decomposable if C satisfies these properties). Then, the reduction from $\mathsf{SHAP}_{\Pi.}(\cdot,\cdot,\cdot)$ to $\mathsf{H}_{\Pi.}(\cdot,\cdot,\cdot)$ from Section 3.1.1 still works, for instance the term R from Equation (4) becomes

$$\begin{aligned} R &= \sum_{\substack{S \subseteq X \setminus \{x\} \\ |S|=k}} \mathbb{E}_{\mathbf{e}' \sim \Pi_p}[C(\mathbf{e}') \mid \mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)] \\ &= \sum_{v \in \mathsf{dom}(X)} \mathrm{p}_x(v) \cdot \mathsf{H}_{\Pi_{p_{|X\setminus\{x\}}}}(C_{x=v}, \mathbf{e}_{|X\setminus\{x\}}, k). \end{aligned}$$

We now look at the computation of $\mathsf{H}_{\Pi.}(\cdot,\cdot,\cdot)$ and inspect what changes in Lemma 4. We can rewrite to fan-in 2 and smooth the circuit in the same way,[6] and the quantities α_g^ℓ are also defined in the same way. The relations that we used to compute the values α_g^ℓ do not change, except the ones for variable gates: for a variable gate g labeled with $x = v$, Equations (8) and (9) become, respectively, $\alpha_g^0 = \mathrm{p}_x(v)$ and

$$\alpha_g^1 = \begin{cases} 1 & \text{if } \mathbf{e}(x) = v \\ 0 & \text{otherwise} \end{cases}.$$

■

This in particular allows us to prove that the SHAP-score can be computed in polynomial time for (not necessarily binary) decision trees, or for variants of OBDDs/FBDDs that use non-binary features, as deterministic and decomposable non-binary Boolean circuits generalize them.

5. Limits on the Tractable Computation of the SHAP-Score

We have shown in the previous sections that the SHAP-score can be computed in polynomial time for deterministic and decomposable circuits under product distributions. A natural question, then, is whether both determinism and decomposability are necessary for this positive result to hold. In this section we show that this is the case, at least under standard complexity assumptions, and even when we consider the uniform distribution. Recall that we are now back to considering Boolean classifiers. Formally, we prove the following:

Theorem 6 *The following problems are* #P*-hard.*

1. *Given as input a decomposable (but not necessarily deterministic) Boolean circuit C over a set of features X, an entity $\mathbf{e} : X \to \{0,1\}$, and a feature $x \in X$, compute the value $\mathsf{SHAP}_{\mathcal{U}}(C, \mathbf{e}, x)$.*

6. For smoothing, we use the circuits $d_S \coloneqq \bigwedge_{x \in S}(\bigvee_{v \in \mathsf{dom}(x)} x = v)$.

2. *Given as input a deterministic (but not necessarily decomposable) Boolean circuit C over a set of features X, an entity $\mathbf{e} : X \to \{0,1\}$, and a feature $x \in X$, compute the value $\mathsf{SHAP}_{\mathcal{U}}(C, \mathbf{e}, x)$.*

Intuitively, for the first item, this comes from the fact that an arbitrary Boolean circuit can always be transformed in polynomial time (in fact in linear time) into an equivalent decomposable (but not necessarily deterministic) Boolean circuit, simply by applying De Morgan's laws to eliminate all $\wedge$-gates. Hence the problem on those circuits it at least as hard as on unrestricted Boolean circuits; and the argument is similar for the second item. Therefore, to show Theorem 6, it is good enough to prove that the problem is indeed intractable over unrestricted Boolean circuits. We now prove these claims formally.

We start by showing that there is a general polynomial-time reduction from the problem of computing the number of entities that satisfy M, for M an arbitrary Boolean classifier, to the problem of computing the SHAP-score over M under the uniform distribution. This holds under the mild condition that $M(\mathbf{e})$ can be computed in polynomial time for an input entity $\mathbf{e}$, which is satisfied for all the Boolean circuits and binary decision diagrams classes considered in this paper. The proof of this result follows from well-known properties of Shapley values, and a closely related result can be found as Theorem 5.1 in (Bertossi et al., 2020).

Lemma 7 *Let M be a Boolean classifier over a set of features X, and let $\#\mathsf{SAT}(M) := |\{\mathbf{e} \in \mathsf{ent}(X) \mid M(\mathbf{e}) = 1\}|$. Then for every $\mathbf{e} \in \mathsf{ent}(X)$ we have:*

$$\#\mathsf{SAT}(M) \;=\; 2^{|X|}\Big(M(\mathbf{e}) - \sum_{x \in X} \mathsf{SHAP}_{\mathcal{U}}(M, \mathbf{e}, x)\Big).$$

Proof The validity of this equation will be a consequence of the following property of the SHAP-score: for every Boolean classifier M over X, entity $\mathbf{e} \in \mathsf{ent}(X)$ and feature $x \in X$, it holds that

$$\sum_{x \in X} \mathsf{SHAP}_{\mathcal{U}}(M, \mathbf{e}, x) \;=\; \phi_{\mathcal{U}}(M, \mathbf{e}, X) - \phi_{\mathcal{U}}(M, \mathbf{e}, \emptyset). \tag{15}$$

This property is often called the *efficiency* property of the Shapley value. Although this is folklore, we prove Equation (15) here for the reader's convenience. Recall from Equation (2) that the SHAP-score can be written as

$$\mathsf{SHAP}_{\mathcal{U}}(M, \mathbf{e}, x) \;= \frac{1}{|X|!} \sum_{\pi \in \Pi(X)} \big(\phi_{\mathcal{U}}(M, \mathbf{e}, S^x_\pi \cup \{x\}) - \phi_{\mathcal{U}}(M, \mathbf{e}, S^x_\pi)\big).$$

Hence, we have that

$$\begin{aligned}
\sum_{x \in X} \mathsf{SHAP}_{\mathcal{U}}(M, \mathbf{e}, x) \;&=\; \frac{1}{|X|!} \sum_{x \in X} \sum_{\pi \in \Pi(X)} \big(\phi_{\mathcal{U}}(M, \mathbf{e}, S^x_\pi \cup \{x\}) - \phi_{\mathcal{U}}(M, \mathbf{e}, S^x_\pi)\big) \\
&=\; \frac{1}{|X|!} \sum_{\pi \in \Pi(X)} \sum_{x \in X} \big(\phi_{\mathcal{U}}(M, \mathbf{e}, S^x_\pi \cup \{x\}) - \phi_{\mathcal{U}}(M, \mathbf{e}, S^x_\pi)\big)
\end{aligned}$$

$$= \frac{1}{|X|!} \sum_{\pi \in \Pi(X)} \big(\phi_{\mathcal{U}}(M, \mathbf{e}, X) - \phi_{\mathcal{U}}(M, \mathbf{e}, \emptyset)\big)$$
$$= \phi_{\mathcal{U}}(M, \mathbf{e}, X) - \phi_{\mathcal{U}}(M, \mathbf{e}, \emptyset),$$

where the second to last equality is obtained by noticing that the inner sum is a telescoping sum. This establishes Equation (15). Now, by definition of $\phi_{\mathcal{U}}(\cdot, \cdot, \cdot)$ we have that $\phi_{\mathcal{U}}(M, \mathbf{e}, X) = M(\mathbf{e})$ and $\phi_{\mathcal{U}}(M, \mathbf{e}, \emptyset) = \frac{1}{2^{|X|}} \sum_{\mathbf{e}' \in \mathsf{ent}(X)} M(\mathbf{e}')$, so we obtain

$$\sum_{x \in X} \mathsf{SHAP}_{\mathcal{U}}(M, \mathbf{e}, x) \;=\; M(\mathbf{e}) - \frac{\#\mathsf{SAT}(M)}{2^{|X|}},$$

thus proving Lemma 7. ■

We stress out here that this result holds for *any* Boolean classifier and is not restricted to the classes of Boolean circuits that we consider in this paper.

Theorem 6 can then easily be deduced from Lemma 7 and the following two facts: (a) counting the number of satisfying assignments of an arbitrary Boolean circuit is a #P-hard problem (Provan and Ball, 1983); and (b) every Boolean circuit can be transformed in linear time into a Boolean circuit that is deterministic (resp, decomposable), simply by using De Morgan's laws to get rid of the $\vee$-gates (resp., $\wedge$-gates). For instance, the reduction to prove the first item of Theorem 6 is as follows: on input an arbitrary Boolean circuit C, use De Morgan's laws to remove all $\wedge$-gates, thus obtaining an equivalent circuit C' which is vacuously decomposable (since it does not have any $\wedge$-gates), and then use the oracle to computing SHAP-scores together with Lemma 7 with an arbitrary entity (for instance, the one that assigns zero to all features) in order to compute $\#\mathsf{SAT}(C)$.

6. Non-Approximability of the SHAP-Score

We now study the approximability of the SHAP-score. As we have shown in the previous section with Lemma 7, computing the SHAP-score is generally intractable for classes of models for which model counting is intractable. Yet, it could be the case that one can efficiently approximate the SHAP-score, just like in some cases one can efficiently approximate the number of models of a formula even though computing this quantity exactly is intractable. For instance, model counting of DNF formulas is #P-hard (Provan and Ball, 1983) but admits a *Fully Polynomial-time Randomized Approximation Scheme*, or FPRAS (Karp et al., 1989). Unfortunately, as we show next, intractability of computing the SHAP-score continues to hold when one considers approximability, and this even for very restricted kinds of Boolean classifiers and under the uniform probability distribution.

To simplify the notation in this section, we will drop the subscript $\mathcal{U}$ for the uniform distribution, and write $\mathsf{SHAP}(\cdot, \cdot, \cdot)$ instead of $\mathsf{SHAP}_{\mathcal{U}}(\cdot, \cdot, \cdot)$, and $\phi(\cdot, \cdot, \cdot)$ instead of $\phi_{\mathcal{U}}(\cdot, \cdot, \cdot)$. Let $\mathcal{C}$ be a class of Boolean classifiers and $\varepsilon \in (0, 1)$. We say that the problem of computing the SHAP score for $\mathcal{C}$ admits an ε *polynomial-time randomized approximation* (ε-PRA), if there exists a randomized algorithm $\mathcal{A}$ satisfying the following conditions. For every Boolean classifier $M \in \mathcal{C}$ over a set of features X, entity $\mathbf{e}$ over X, and feature $x \in X$, it

holds that:

$$\Pr\left(|\mathcal{A}(M,\mathbf{e},x) - \mathsf{SHAP}(M,\mathbf{e},x)| \leq \varepsilon \cdot |\mathsf{SHAP}(M,\mathbf{e},x)|\right) \quad \geq \quad \frac{3}{4}.$$

Moreover, there exists a polynomial $p(\cdot)$ such that $\mathcal{A}(M,\mathbf{e},x)$ works in time $O(p(\|M\| + \|\mathbf{e}\|))$, where $\|M\|$ and $\|\mathbf{e}\|$ are the sizes of M and $\mathbf{e}$ represented as input strings, respectively. We recall that the notion of FPRAS mentioned before is defined as the concept of PRA, but imposing the stronger requirements that ε is part of the input and the algorithm is polynomial in $\frac{1}{\varepsilon}$ as well.

We start by presenting a simple proof that for every $\varepsilon \in (0,1)$, the problem of computing the SHAP score for Boolean classifiers given as DNF formulas does not admit an ε-PRA, unless NP = RP.[7]

Proposition 8 *For every $\varepsilon \in (0,1)$, the problem of computing the SHAP score for Boolean classifiers given as DNF formulas does not admit an ε-PRA, unless* NP = RP*. This result holds even if we restrict to the uniform distributions on the entities.*

Proof We first recall the following fact: if a function f admits an ε-PRA, then the problem of determining, given a string x, whether $f(x) = 0$ is in BPP (although this is folklore, we provide a proof in Appendix B). We use this to prove that if we could approximate the SHAP-score over DNF formulas, then we could solve the validity problem over DNF formulas in BPP. Recall that the validity problem over DNF formulas is the decision problem that, given as input a DNF formula φ, accepts if all valuations satisfy φ, and rejects otherwise (in other words, it rejects if $\neg\varphi$ is satisfiable and accepts otherwise). Since this problem is coNP-complete and BPP is closed under complement, this would imply that NP $\subseteq$ BPP, and thus that NP = RP (Ko, 1982). Let φ be a DNF formula over a set of variables X. We consider the DNF formula $\varphi' \coloneqq \varphi \vee x$ with $x \notin X$, and the uniform probability distribution over $\mathsf{ent}(X \cup \{x\})$. Let $\mathbf{e}$ be an arbitrary entity over $X \cup \{x\}$ such that $\mathbf{e}(x) = 1$. We show that φ is valid if and only if $\mathsf{SHAP}(\varphi', \mathbf{e}, x) = 0$, which, by the previous remarks, is good enough to conclude the proof. By definition we have

$$\mathsf{SHAP}(\varphi', \mathbf{e}, x) \;=\; \sum_{S \subseteq X} \frac{|S|!\,(|X| - |S|)!}{(|X|+1)!} \left(\phi(\varphi', \mathbf{e}, S \cup \{x\}) - \phi(\varphi', \mathbf{e}, S)\right).$$

Observe that for each $S \subseteq X$, it holds that $\phi(\varphi', \mathbf{e}, S \cup \{x\}) = 1$ (given the definition of φ' and the fact that $\mathbf{e}(x) = 1$), and that $0 \leq \phi(\varphi', \mathbf{e}, S) \leq 1$. Now, if φ is valid, then it is clear that $\phi(\varphi', \mathbf{e}, S) = 1$ for every $S \subseteq X$, so that indeed $\mathsf{SHAP}(\varphi', \mathbf{e}, x) = 0$. Assume now that φ is not valid. By what preceded, it is good enough to show that for some $S \subseteq X$ we have $\phi(\varphi', \mathbf{e}, S) < 1$. But this clearly holds for $S = \emptyset$, because φ is not valid and all entities have the same probability (so that no entity has probability zero). ■

Hence, this result already establishes an important difference with model counting: for the case of DNF formulas, the model counting problem admits an FPRAS (Karp et al., 1989) and, thus, an ε-PRA for every $\varepsilon \in (0,1)$.

7. Recall that it is widely believed that RP is properly contained in NP, as mentioned in Section 2.

It is important to mention that the proof of Proposition 8 uses, in a crucial way, the fact that the validity problem for DNF formulas is intractable. We prove next a strong negative result, which establishes that not even for *positive* DNF formulas – for which the validity problem is trivial – it is possible to obtain an ε-PRA for computing the SHAP-score. Let $\varphi = D_1 \vee D_2 \vee \cdots \vee D_k$ be a formula in DNF, that is, each formula D_i is a conjunction of positive literals (propositional variables) and negative literals (negations of propositional variables). Then φ is said to be in POS-DNF if each formula D_i is a conjunction of positive literals, (hence, φ is a monotone formula), and φ is said to be in 2-POS-DNF if φ is in POS-DNF and each formula D_i contains at most two (positive) literals. Our main result of this section is the following.

Theorem 9 *For every $\varepsilon \in (0,1)$, the problem of computing the* SHAP *score for Boolean classifiers given as 2-POS-DNF formulas does not admit an ε-PRA, unless* NP = RP*. This result holds even if we restrict to the uniform distributions on the entities.*

Furthermore, we can show that the same intractability result also holds for closely related classes of Boolean classifiers. As before, let $\varphi = D_1 \vee D_2 \vee \cdots \vee D_k$ be a formula in DNF. Then φ is in NEG-DNF if each formula D_i is a conjunction of negative literals, and φ is in 2-NEG-DNF if it is in NEG-DNF and each formula D_i is a conjunction of at most two (negative) literals. We define similarly formulas in 2-POS-CNF and in 2-NEG-CNF. We then obtain the following result.

Corollary 10 *Let $\mathcal{C} \in \{$2-POS-DNF, 2-NEG-DNF, 2-POS-CNF, 2-NEG-CNF$\}$. Then for every $\varepsilon \in (0,1)$, the problem of computing the* SHAP *score for Boolean classifiers given as formulas in $\mathcal{C}$ does not admit an ε-PRA, unless* NP = RP*. This result holds even if we restrict to the uniform distributions on the entities.*

It should be noticed that 2-NEG-CNF formulas are special cases of HORN-SAT formulas, so that the previous result also applies for HORN-SAT.

We prove Theorem 9 in Section 6.1, and then show how the proof can be adapted to prove Corollary 10 in Section 6.2

6.1 Proof of Theorem 9

We will use some results and techniques related to approximating cliques in graphs. More specifically, all graphs $G = (N, E)$ considered in this proof are assumed to be undirected and loop-free (that is, edges of the form (a, a) are not allowed). Besides, we assume that each graph $G = (N, E)$ satisfies the following condition:

(A) there exist at least two isolated nodes in G, that is, two distinct nodes $a, b \in N$ such that $(a, c) \notin E$ and $(b, c) \notin E$ for every node $c \in N$.

The assumption that condition (A) is satisfied will allow us to simplify some calculations. Besides, it is clear that condition (A) can be checked in polynomial time.

We define the problem GapClique as follows. The input of GapClique is a graph $G = (N, E)$ and a number $m \leq |N|$, and its output is:

- **yes** if G contains a clique with m nodes,

- **no** if every clique of G contains at most $\lfloor \frac{m}{3} \rfloor$ nodes.

From the PCP theorem and its applications to hardness of approximation (Feige et al., 1996; Arora and Safra, 1998; Arora et al., 1998), it is known that GapClique is NP-hard (in fact even if $\frac{1}{3}$ is replaced by any $\delta \in (0,1)$). In other words, there exists a polynomial-time reduction that takes as input a Boolean formula φ and that outputs a graph G and an integer m such that (1) if φ is satisfiable then G contains a clique with m nodes; and (2) if φ is not satisfiable then every clique of G contains at most $\lfloor \frac{m}{3} \rfloor$ nodes.

The idea of our proof of Theorem 9 is then to show that if the problem of computing the SHAP score for Boolean classifiers given as 2-POS-DNF formulas admits an ε-PRA, then there exists a BPP algorithm for GapClique. Hence, we would conclude that NP $\subseteq$ BPP, which in turn implies that NP = RP (Ko, 1982). The proof is technical, and it is divided into five modular parts, highlighted in bold in the remaining of this section.

The SHAP-score of a Boolean classifier of the form $M \vee x$. Given a set of features X, a Boolean classifier M over X, and an $S \subseteq X$, define

$$\#\mathsf{SAT}(M,S) \coloneqq |\{\mathbf{e} \in \mathsf{ent}(X) \mid M(\mathbf{e}) = 1 \text{ and } \mathbf{e}(y) = 1 \text{ for every } y \in S\}|.$$

We can relate the SHAP-score of a Boolean classifier of the form $M \vee x$ to the quantities $\#\mathsf{SAT}(\neg M, S)$ as follows:

Lemma 11 *Let X be a set of features, $n = |X|$, $x \in X$, M be a Boolean classifier over $X \setminus \{x\}$, and $\mathbf{1}$ be the entity over X such that $\mathbf{1}(y) = 1$ for every $y \in X$. Then*

$$\mathsf{SHAP}(M \vee x, \mathbf{1}, x) \;=\; \sum_{k=0}^{n-1} \frac{k!(n-k-1)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq X \setminus \{x\} \,:\, |S| = k} \#\mathsf{SAT}(\neg M, S).$$

Proof Let M' be a Boolean classifier over X. Given $S \subseteq X \setminus \{x\}$, we have that:

$$\begin{aligned}
\phi(M', \mathbf{1}, S \cup \{x\}) &= \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}, S \cup \{x\})} \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \cdot M'(\mathbf{e}) \\
&= \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M'_{+x}(\mathbf{e}') \\
&= \frac{2}{2^{|X \setminus S|}} \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M'_{+x}(\mathbf{e}').
\end{aligned}$$

Besides, we have that:

$$\begin{aligned}
\phi(M', \mathbf{1}, S) &= \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}, S)} \frac{1}{2^{|X \setminus S|}} \cdot M'(\mathbf{e}) \\
&= \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}, S) \,:\, \mathbf{e}(x) = 0} \frac{1}{2^{|X \setminus S|}} \cdot M'(\mathbf{e}) \;+\; \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}, S) \,:\, \mathbf{e}(x) = 1} \frac{1}{2^{|X \setminus S|}} \cdot M'(\mathbf{e}) \\
&= \frac{1}{2^{|X \setminus S|}} \cdot \left(\sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M'_{-x}(\mathbf{e}') \;+\; \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M'_{+x}(\mathbf{e}') \right).
\end{aligned}$$

Therefore, we conclude that:

$$\phi(M', \mathbf{1}, S \cup \{x\}) - \phi(M', \mathbf{1}, S) \;=\; \frac{1}{2^{|X \setminus S|}} \cdot \Big(\sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M'_{+x}(\mathbf{e}) - \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M'_{-x}(\mathbf{e}) \Big).$$

Considering this equation with $M' \coloneqq M \vee x$, we deduce that:

$$\begin{aligned}
&\phi(M \vee x, \mathbf{1}, S \cup \{x\}) - \phi(M \vee x, \mathbf{1}, S) \;= \\
&\qquad \frac{1}{2^{|X \setminus S|}} \cdot \Big(\sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} (M \vee x)_{+x}(\mathbf{e}) - \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} (M \vee x)_{-x}(\mathbf{e}) \Big) \;= \\
&\qquad \frac{1}{2^{|X \setminus S|}} \cdot \Big(\sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} 1 - \sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} M(\mathbf{e}) \Big) \;= \\
&\qquad \frac{1}{2^{|X \setminus S|}} \cdot \Big(\sum_{\mathbf{e} \in \mathsf{cw}(\mathbf{1}_{|X \setminus \{x\}}, S)} (1 - M(\mathbf{e})) \Big) \;= \\
&\qquad \frac{1}{2^{|X \setminus S|}} \cdot \#\mathsf{SAT}(\neg M, S).
\end{aligned}$$

By considering the definition of the SHAP score, we obtain that:

$$\begin{aligned}
&\mathsf{SHAP}(M \vee x, \mathbf{1}, x) \;= \\
&\qquad \sum_{S \subseteq X \setminus \{x\}} \frac{|S|!(|X| - |S| - 1)!}{|X|!} \cdot \Big(\phi(M \vee x, \mathbf{1}, S \cup \{x\}) - \phi(M \vee x, \mathbf{1}, S) \Big) \;= \\
&\qquad \sum_{k=0}^{n-1} \sum_{S \subseteq X \setminus \{x\} \,:\, |S| = k} \frac{|S|!(|X| - |S| - 1)!}{|X|!} \cdot \frac{1}{2^{|X \setminus S|}} \cdot \#\mathsf{SAT}(\neg M, S) \;= \\
&\qquad \sum_{k=0}^{n-1} \frac{k!(n-k-1)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq X \setminus \{x\} \,:\, |S| = k} \#\mathsf{SAT}(\neg M, S).
\end{aligned}$$

This concludes the proof of Lemma 11. ∎

A 2-POS-DNF formula for cliques. Given a graph $G = (N, E)$ we define the formula $\theta(G)$ in 2-POS-DNF by

$$\theta(G) \coloneqq \bigvee_{\substack{(a,b) \in (N \times N) \\ a \neq b \text{ and } (a,b) \notin E}} a \wedge b. \tag{16}$$

Notice that the set of propositional variables occurring in $\theta(G)$ is the same as the set N of nodes of G, given that G satisfies condition (A). For $S \subseteq N$ we define

$$\#\mathsf{CLIQUE}(G, S) \;\coloneqq\; |\{Y \mid Y \text{ is a clique of } G \text{ and } S \subseteq Y\}|.$$

Using Lemma 11, we can relate the SHAP-score of the 2-POS-DNF formula $\theta(G) \vee x$ to the quantities $\#\mathsf{CLIQUE}(G, S)$ for $S \subseteq N$ as follows.

Lemma 12 *For every graph $G = (N, E)$ and $x \notin N$, letting $n = |N|$ and $\mathbf{1}$ be the entity over $N \cup \{x\}$ such that $\mathbf{1}(y) = 1$ for every $y \in N \cup \{x\}$, we have:*

$$\mathsf{SHAP}(\theta(G) \vee x, \mathbf{1}, x) = \frac{1}{2(n+1)} \sum_{k=0}^{n} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq N \,:\, |S|=k} \#\mathsf{CLIQUE}(G, S).$$

Proof The crucial observation is that the following relation holds for every $S \subseteq N$:

$$\#\mathsf{CLIQUE}(G, S) \quad = \quad \#\mathsf{SAT}(\neg\theta(G), S).$$

Indeed, this can be seen by defining the bijection that to every subset Y of N associates the valuation ν_Y of $\theta(G)$ such that $\nu_Y(x) = 1$ if and only if $x \in Y$, and then checking that [Y is a clique of G with $S \subseteq Y$] if and only if [$\nu_Y \models \neg\theta(G)$ and $\nu_Y(x) = 1$ for all $x \in S$]. Therefore, we have from Lemma 11 that

$$\begin{aligned}
\mathsf{SHAP}(\theta(G) \vee x, \mathbf{1}, x) &= \sum_{k=0}^{n} \frac{k!(n-k)!}{(n+1)!} \cdot \frac{1}{2^{n+1-k}} \sum_{S \subseteq N \,:\, |S|=k} \#\mathsf{SAT}(\neg\theta(G), S) \\
&= \frac{1}{2(n+1)} \sum_{k=0}^{n} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq N \,:\, |S|=k} \#\mathsf{CLIQUE}(G, S).
\end{aligned}$$

■

Working with amplified graphs. In this proof, we consider an amplification technique from (Sinclair, 1993). More precisely, given a graph $G = (N, E)$ and a natural number $r \geq 1$, define the amplified graph $G^r = (N^r, E^r)$ of G as follows. For each $a \in N$, let $N^a := \{a_1, \ldots, a_r\}$ and $N^r := \bigcup_{a \in N} N^a$, and let E^r be the following set of edges:

$$E^r \quad := \quad \{(a_i, a_j) \mid a_i, a_j \in N^a \text{ and } i \neq j\} \cup \{(a_i, b_j) \mid a_i \in N^a, b_j \in N^b \text{ and } (a, b) \in E\}.$$

Thus, the amplified graph G^r is constructed by copying r times each node of G, by connecting for each node of G all of its copies as a clique, and finally by connecting copies a_i, b_j of nodes a, b of G, respectively, whenever a and b are connected in G by an edge. Notice in particular that G^1 is simply G. An example of this construction for $r = 2$ is illustrated in Figure 3.

Let then T be a clique of G^r. We say that T *is a witness* of the clique $\{a \in N \mid N^a \cap T \neq \emptyset\}$ of G. Observe that a clique of G^r witnesses a unique clique of G by definition, but that a clique of G can have multiple witnessing cliques in G^r. Moreover, letting S be a clique of G, we write $\mathsf{Wit}(G^r, S)$ for the set of cliques of G^r that are witnesses of S. We then show the following three properties of G^r:

(i) if S_1, S_2 are distinct cliques of G, then $\mathsf{Wit}(G^r, S_1) \cap \mathsf{Wit}(G^r, S_2) = \emptyset$. Indeed, assuming without loss of generality that $S_1 \not\subseteq S_2$ (the case $S_2 \not\subseteq S_1$ being symmetrical) and letting $a \in S_1 \setminus S_2$, it is clear by that, by definition of being a witness, for any $T \in \mathsf{Wit}(G^r, S_1)$ we must have $N^a \cap T \neq \emptyset$, whereas for any $T \in \mathsf{Wit}(G^r, S_2)$ we have $N^a \cap T = \emptyset$.

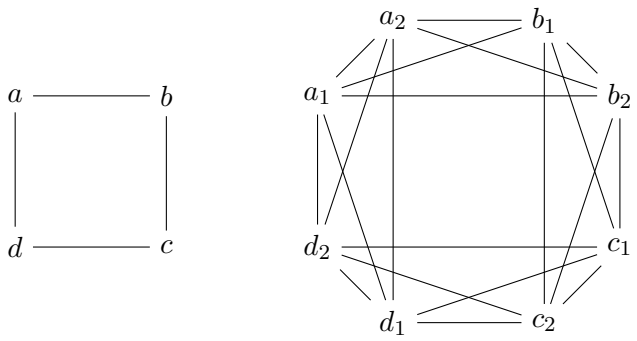


Figure 3: Illustration of the amplification construction for $r = 2$. On the left a graph G, on the right the corresponding graph G^2. We only illustrate the construction for $r = 2$, as for $r \geq 3$ this very quickly becomes unreadable.

(ii) if S is a clique of G, then $|\mathsf{Wit}(G^r, S)| = (2^r - 1)^{|S|}$. Indeed, let $T \in \mathsf{Wit}(G^r, S)$. By definition of being a witness, T cannot contain any node of the form a_i for $a \notin S$. Moreover, for every $a \in S$, by definition of being a witness again, the set $T \cap N^a$ must be a non-empty subset of N^a. Since for every $a \in N$ there are exactly $(2^r - 1)$ non-empty subsets of N^a, this shows that $|\mathsf{Wit}(G^r, S)| \leq (2^r - 1)^{|S|}$. Additionally, any set of the form $\bigcup_{a \in S} S_a$ where each S_a is a non-empty subset of N^a is in fact in $\mathsf{Wit}(G^r, S)$: this is simply because S itself is a clique of S and by definition of the graph G^r. This shows that $|\mathsf{Wit}(G^r, S)| \geq (2^r - 1)^{|S|}$, hence $|\mathsf{Wit}(G^r, S)| = (2^r - 1)^{|S|}$ indeed.

(iii) for every natural number ℓ, if all cliques of G have size at most ℓ, then all cliques of G^r have size at most $\ell \cdot r$. Indeed, for any clique T of G^r, letting S be the clique of S that T witnesses, it is clear that $|T| \leq r \cdot |S|$ (with the equality being reached when $T \cap N^a = N^a$ for every $a \in S$).

We use these properties to prove our next lemma.

Lemma 13 *Let $G = (N, E)$ be a graph, $x \notin N$, $n \coloneqq |N|$, and m, r be natural numbers such that $1 \leq m \leq n$ and $r \geq 1$. Then:*

(a) If every clique of G contains at most $\lfloor \frac{m}{3} \rfloor$ nodes, then

$$\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) \quad \leq \quad \frac{2^{2\lfloor \frac{m}{3} \rfloor r + n}}{2^{n \cdot r + 1}}.$$

(b) If G contains a clique with m nodes, then

$$\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) \quad \geq \quad \frac{1}{(n \cdot r + 1)} \cdot \frac{2^{mr}}{2^{n \cdot r + 1}}.$$

Proof

(a) Given that $|N^r| = n \cdot r$, we have from Lemma 12 that:

$$\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) \;=\; \frac{1}{2(n \cdot r + 1)} \sum_{k=0}^{n \cdot r} \frac{k!(n \cdot r - k)!}{(n \cdot r)!} \cdot \frac{1}{2^{n \cdot r - k}} \sum_{S \subseteq N^r \,:\, |S| = k} \#\mathsf{CLIQUE}(G^r, S).$$

Now, since every clique of G contains at most $\lfloor \frac{m}{3} \rfloor$ nodes, we have by Item (iii) that all cliques of G^r have size at most $\lfloor \frac{m}{3} \rfloor \cdot r$. Hence when $k > \lfloor \frac{m}{3} \rfloor \cdot r$ it holds that $\#\mathsf{CLIQUE}(G^r, S) = 0$ for any $S \subseteq N^r$ with $|S| = k$. Therefore, we have that

$$\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) \;=$$

$$\frac{1}{2(n \cdot r + 1)} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \frac{k!(n \cdot r - k)!}{(n \cdot r)!} \cdot \frac{1}{2^{n \cdot r - k}} \sum_{S \subseteq N^r \,:\, |S| = k} \#\mathsf{CLIQUE}(G^r, S) \;\leq$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \frac{k!(n \cdot r - k)!}{(n \cdot r)!} \sum_{S \subseteq N^r \,:\, |S| = k} \#\mathsf{CLIQUE}(G^r, S) \;=$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \frac{1}{\binom{n \cdot r}{k}} \sum_{S \subseteq N^r \,:\, |S| = k} \#\mathsf{CLIQUE}(G^r, S).$$

Now, defining $\#\mathsf{CLIQUE}(G, \ell) \coloneqq |\{Y \mid Y \text{ is a clique of } G \text{ and } |Y| = \ell\}|$, observe that we have

$$\begin{aligned}
\sum_{S \subseteq N^r \,:\, |S| = k} \#\mathsf{CLIQUE}(G^r, S) \;&=\; \sum_{S \subseteq N^r \,:\, |S| = k} \;\sum_{\substack{Y \text{ clique of } G^r \\ S \subseteq Y}} 1 \\
&=\; \sum_{S \subseteq N^r \,:\, |S| = k} \;\sum_{\ell = k}^{|N^r|} \;\sum_{\substack{Y \text{ clique of } G^r \\ |Y| = \ell \\ S \subseteq Y}} 1 \\
&=\; \sum_{S \subseteq N^r \,:\, |S| = k} \;\sum_{\ell = k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \;\sum_{\substack{Y \text{ clique of } G^r \\ |Y| = \ell \\ S \subseteq Y}} 1 \\
&=\; \sum_{\ell = k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \;\sum_{S \subseteq N^r \,:\, |S| = k} \;\sum_{\substack{Y \text{ clique of } G^r \\ |Y| = \ell \\ S \subseteq Y}} 1 \\
&=\; \sum_{\ell = k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \;\sum_{\substack{Y \text{ clique of } G^r \\ |Y| = \ell}} \;\sum_{S \subseteq Y \,:\, |S| = k} 1
\end{aligned}$$

$$= \sum_{\ell=k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \sum_{\substack{Y \text{ clique of } G^r \\ |Y|=\ell}} \binom{|Y|}{k}$$

$$= \sum_{\ell=k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \binom{\ell}{k} \sum_{\substack{Y \text{ clique of } G^r \\ |Y|=\ell}} 1$$

$$= \sum_{\ell=k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \binom{\ell}{k} \#\mathsf{CLIQUE}(G^r, \ell),$$

where the third equality is simply because all cliques of G^r have size at most $\lfloor \frac{m}{3} \rfloor \cdot r$, again by Item (iii). Hence, we have that

$$\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) \leq$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \frac{1}{\binom{n \cdot r}{k}} \sum_{\ell=k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \binom{\ell}{k} \#\mathsf{CLIQUE}(G^r, \ell) \leq$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \frac{1}{\binom{n \cdot r}{k}} \sum_{\ell=k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \binom{n \cdot r}{k} \#\mathsf{CLIQUE}(G^r, \ell) =$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \sum_{\ell=k}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell) \leq$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{k=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \sum_{\ell=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell) \leq$$

$$\frac{(\lfloor \frac{m}{3} \rfloor \cdot r + 1)}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{\ell=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell) \leq$$

$$\frac{n \cdot r + 1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{\ell=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell) =$$

$$\frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{\ell=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell).$$

But notice that $\sum_{\ell=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell)$ is simply the total number of cliques of G^r. Given a clique S of G with ℓ elements, remember that by Item (ii) we have that $|\mathsf{Wit}(G^r, S)| = (2^r - 1)^\ell$. Hence, given that each clique of G has at most $\lfloor \frac{m}{3} \rfloor$ nodes, that G has at most $\binom{n}{\ell}$ cliques with ℓ nodes, and that every clique of G^r witnesses a unique clique of G, we have that the total number of cliques of G^r is bounded as follows:

$$\sum_{\ell=0}^{\lfloor \frac{m}{3} \rfloor \cdot r} \#\mathsf{CLIQUE}(G^r, \ell) \leq \sum_{\ell'=0}^{\lfloor \frac{m}{3} \rfloor} \binom{n}{\ell'} (2^r - 1)^{\ell'}.$$

Therefore, we conclude that:

$$\begin{aligned}
\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) &\leq \frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{\ell'=0}^{\lfloor \frac{m}{3} \rfloor} \binom{n}{\ell'} (2^r - 1)^{\ell'} \\
&\leq \frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \sum_{\ell'=0}^{\lfloor \frac{m}{3} \rfloor} \binom{n}{\ell'} (2^r - 1)^{\lfloor \frac{m}{3} \rfloor} \\
&= \frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \cdot (2^r - 1)^{\lfloor \frac{m}{3} \rfloor} \sum_{\ell'=0}^{\lfloor \frac{m}{3} \rfloor} \binom{n}{\ell'} \\
&\leq \frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \cdot (2^r - 1)^{\lfloor \frac{m}{3} \rfloor} \sum_{\ell'=0}^{n} \binom{n}{\ell'} \\
&= \frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \cdot (2^r - 1)^{\lfloor \frac{m}{3} \rfloor} \cdot 2^n \\
&\leq \frac{1}{2} \cdot \frac{1}{2^{n \cdot r - \lfloor \frac{m}{3} \rfloor r}} \cdot 2^{\lfloor \frac{m}{3} \rfloor r} \cdot 2^n \\
&= \frac{2^{2\lfloor \frac{m}{3} \rfloor r + n}}{2^{n \cdot r + 1}}.
\end{aligned}$$

(b) First, we claim that G^r contains at least 2^{mr} cliques. Indeed, let S be a clique with m nodes that G contains (by hypothesis). Every subset S' of S is also a clique of G. Now, by Item (i) and Item (ii), we have that the number of witnesses of all the subcliques of S is equal to

$$\begin{aligned}
\left| \bigcup_{S' \subseteq S} \mathsf{Wit}(G^r, S') \right| &= \sum_{S' \subseteq S} |\mathsf{Wit}(G^r, S')| \\
&= \sum_{\ell=0}^{m} \binom{m}{\ell} (2^r - 1)^{\ell} \\
&= \sum_{\ell=0}^{m} \binom{m}{\ell} (2^r - 1)^{\ell} 1^{m-\ell} \\
&= (2^r - 1 + 1)^m \\
&= 2^{mr}.
\end{aligned}$$

All terms in the summation of $\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x)$ are non-negative. Hence, we have that:

$$\begin{aligned}
&\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) = \\
&\frac{1}{2(n \cdot r + 1)} \sum_{k=0}^{n \cdot r} \frac{k!(n \cdot r - k)!}{n \cdot r!} \cdot \frac{1}{2^{n \cdot r - k}} \sum_{S \subseteq N^r : |S| = k} \#\mathsf{CLIQUE}(G^r, S) \geq \\
&\frac{1}{2(n \cdot r + 1)} \cdot \frac{0!(n \cdot r)!}{(n \cdot r)!} \cdot \frac{1}{2^{n \cdot r}} \sum_{S \subseteq N^r : |S| = 0} \#\mathsf{CLIQUE}(G^r, S) =
\end{aligned}$$

$$\frac{1}{2(n \cdot r + 1)} \cdot \frac{1}{2^{n \cdot r}} \cdot \#\mathsf{CLIQUE}(G^r, \emptyset).$$

But $\#\mathsf{CLIQUE}(G^r, \emptyset)$ is the total number of cliques of G^r, which we have just proved to be greater than or equal to 2^{mr}. Therefore

$$\mathsf{SHAP}(\theta(G^r) \vee x, \mathbf{1}, x) \quad \geq \quad \frac{1}{(n \cdot r + 1)} \cdot \frac{2^{mr}}{2^{n \cdot r + 1}},$$

which concludes the proof of Lemma 13.

■

A technical lemma. Last, we will need the following technical lemma.

Lemma 14 *For every $\varepsilon \in (0, 1)$, there exists $n_\varepsilon \in \mathbb{N}$ such that for every $n, m \in \mathbb{N}$ with $n \geq n_\varepsilon$ and $m \geq 1$, it holds that:*

$$(1 + \varepsilon) \cdot 2^{2\lfloor \frac{m}{3} \rfloor n^2 + n} \quad < \quad (1 - \varepsilon) \cdot \frac{2^{mn^2}}{(n^3 + 1)}.$$

The proof is straightforward and can be found in Appendix C.1.

Putting it all together. We now have all the necessary ingredients to prove Theorem 9. Assume that the problem of computing the SHAP score for Boolean classifiers given as 2-POS-DNF formulas admits an ε-PRA, for some fixed $\varepsilon \in (0, 1)$, that we will denote by $\mathcal{A}$. By using $\mathcal{A}$, we define the following BPP algorithm $\mathcal{B}$ for $\mathsf{GapClique}$. Let $G = (N, E)$ be a graph with $n = |N|$ and $m \in \{0, \ldots, n\}$. Then $\mathcal{B}(G, m)$ performs the following steps:

1. If $m = 0$, then return **yes**.
2. Let n_ε be the constant in Lemma 14. If $n \leq n_\varepsilon$, then by performing an exhaustive search, check whether G contains a clique with m nodes or if every clique of G contains at most $\lfloor \frac{m}{3} \rfloor$ nodes. In the former case return **yes**, in the latter case return **no**.
3. Construct the formula $\theta(G^{n^2}) \vee x$, where x is a fresh feature (not occurring in N^{n^2}). Notice that $\theta(G^{n^2}) \vee x$ is a formula in 2-POS-DNF.
4. Use algorithm $\mathcal{A}$ to compute $s \coloneqq \mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)$.
5. If
$$s \quad > \quad (1 + \varepsilon) \cdot \frac{2^{2\lfloor \frac{m}{3} \rfloor n^2 + n}}{2^{n^3 + 1}},$$
then return **yes**; otherwise return **no**.

Algorithm $\mathcal{B}$ works in polynomial time since n_ε is a fixed natural number (given that ε is a fixed value in $(0, 1)$), amplified graph G^{n^2} can be computed in polynomial time from G, formula $\theta(G^{n^2}) \vee x$ can be constructed in polynomial time from G^{n^2}, algorithm $\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)$ works in time $p(\|\theta(G^{n^2}) \vee x\| + \|\mathbf{1}\|)$ for a polynomial $p(\cdot)$, and bound $(1 + \varepsilon) \cdot \frac{2^{2\lfloor \frac{m}{3} \rfloor n^2 + n}}{2^{n^3 + 1}}$ can be computed in polynomial time in n. Therefore, to conclude the proof we need to show that the error probability of algorithm $\mathcal{B}$ is bounded by $\frac{1}{4}$.

- Assume that every clique of G contains at most $\lfloor \frac{m}{3} \rfloor$ nodes. Then the error probability of algorithm $\mathcal{B}$ is equal to $\Pr(\mathcal{B}(G,m)$ returns **yes**$)$. By using the definition of $\mathcal{B}$ and Lemma 13 (a) with $r = n^2$, we obtain that:

$$\begin{aligned}
&\Pr(\mathcal{B}(G,m) \text{ returns } \mathbf{yes}) \;=\\
&\Pr\left(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) > (1+\varepsilon) \cdot \frac{2^{2\lfloor \frac{m}{3} \rfloor n^2 + n}}{2^{n^3+1}}\right) \;\leq\\
&\Pr\left(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) > (1+\varepsilon) \cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\right) \;\leq\\
&\Pr\Big(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) > (1+\varepsilon) \cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \vee\\
&\qquad \mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) < (1-\varepsilon) \cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\Big) \;=\\
&1 - \Pr\Big(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \leq (1+\varepsilon) \cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \wedge\\
&\qquad \mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \geq (1-\varepsilon) \cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\Big) \;=\\
&1 - \Pr\left(\left|\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) - \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\right| \;\leq\right.\\
&\qquad\qquad \left.\varepsilon \cdot \left|\mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\right|\right) \;\leq\\
&1 - \frac{3}{4} \;=\; \frac{1}{4},
\end{aligned}$$

  where in the last step we use the fact that $\mathcal{A}$ is an ε-PRA for the problem of computing the SHAP score for Boolean classifiers given as 2-POS-DNF formulas, and the fact that $\mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \geq 0$ (as can be seen from Lemma 12).

- Assume now that G contains a clique with m nodes. We can assume that $n > n_\varepsilon$, since otherwise we know that $\mathcal{B}$ returns the correct answer in Step 2. Given that $n > n_\varepsilon$, we know from Lemma 14 that:

$$(1+\varepsilon) \cdot \frac{2^{2\lfloor \frac{m}{3} \rfloor n^2 + n}}{2^{n^3+1}} \;<\; (1-\varepsilon) \cdot \frac{1}{(n^3+1)} \cdot \frac{2^{mn^2}}{2^{n^3+1}}. \tag{17}$$

  The error probability of algorithm $\mathcal{B}$ is equal to $\Pr(\mathcal{B}(G,m)$ returns **no**$)$. By using the definition of $\mathcal{B}$, Lemma 13 (b) with $r = n^2$, and inequality (17), we obtain that:

$$\begin{aligned}
&\Pr(\mathcal{B}(G,m) \text{ returns } \mathbf{no}) \;=\\
&\Pr\left(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \leq (1+\varepsilon) \cdot \frac{2^{2\lfloor \frac{m}{3} \rfloor n^2 + n}}{2^{n^3+1}}\right) \;\leq\\
&\Pr\left(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) < (1-\varepsilon) \cdot \frac{1}{(n^3+1)} \cdot \frac{2^{mn^2}}{2^{n^3+1}}\right) \;\leq\\
&\Pr\left(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) < (1-\varepsilon) \cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\right) \;\leq
\end{aligned}$$

$$\begin{aligned}
&\Pr\Big(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) < (1-\varepsilon)\cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \vee \\
&\qquad \mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) > (1+\varepsilon)\cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\Big) \;=\\
&1 - \Pr\Big(\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \geq (1-\varepsilon)\cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \wedge \\
&\qquad \mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) \leq (1+\varepsilon)\cdot \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\Big) \;=\\
&1 - \Pr\Big(\Big|\mathcal{A}(\theta(G^{n^2}) \vee x, \mathbf{1}, x) - \mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\Big| \;\leq \\
&\qquad\qquad \varepsilon \cdot \Big|\mathsf{SHAP}(\theta(G^{n^2}) \vee x, \mathbf{1}, x)\Big|\Big) \;\leq \\
&1 - \frac{3}{4} \;=\; \frac{1}{4},
\end{aligned}$$

where the last step is obtained as in the previous case. This concludes the proof of Theorem 9.

6.2 Proof of Corollary 10

First, we explain how the result for 2-NEG-DNF can be obtained from the result for 2-POS-DNF (Theorem 9). For an entity $\mathbf{e}$ over a set of features X, let us write $\mathsf{InvPol}(\mathbf{e})$ the entity defined by $\mathsf{InvPol}(\mathbf{e})(x) \coloneqq 1 - \mathbf{e}(x)$ for every $x \in X$. For a Boolean classifier M over variables X, we write $\mathsf{InvPol}(M)$ the Boolean classifier defined by $\mathsf{InvPol}(M)(\mathbf{e}) \coloneqq M(\mathsf{InvPol}(\mathbf{e}))$; in other words, we have changed the polarity of all the features. Then the following holds:

Lemma 15 *Let M be a Boolean classifier over X, $\mathbf{e}$ an entity over X and $x \in X$. We have*

$$\mathsf{SHAP}(M, \mathbf{e}, x) \;=\; \mathsf{SHAP}(\mathsf{InvPol}(M), \mathsf{InvPol}(\mathbf{e}), x).$$

Proof For every $S \subseteq X$, we have that

$$\begin{aligned}
\phi(\mathsf{InvPol}(M), \mathsf{InvPol}(\mathbf{e}), S) &= \sum_{\mathbf{e}' \in \mathsf{cw}(\mathsf{InvPol}(\mathbf{e}), S)} \frac{1}{2^{|X\setminus S|}} \mathsf{InvPol}(M)(\mathbf{e}') \\
&= \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)} \frac{1}{2^{|X\setminus S|}} \mathsf{InvPol}(M)(\mathsf{InvPol}(\mathbf{e}')) \\
&= \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)} \frac{1}{2^{|X\setminus S|}} M(\mathsf{InvPol}(\mathsf{InvPol}(\mathbf{e}'))) \\
&= \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S)} \frac{1}{2^{|X\setminus S|}} M(\mathbf{e}') \\
&= \phi(M, \mathbf{e}, S),
\end{aligned}$$

which proves the lemma. ■

But it is clear that if M is a Boolean classifier defined with a 2-POS-DNF formula φ, then $\mathsf{InvPol}(M)$ is the Boolean classifier defined by the formula obtained from φ by replacing every occurrence of a variable x by the literal $\neg x$; that is, a formula in 2-NEG-DNF. Combining Theorem 9 with Lemma 15 then establishes Corollary 10 for the case of 2-NEG-DNF. The same argument shows that the results for 2-POS-CNF and 2-NEG-CNF are equivalent, so all that remains to do in this section is to prove the result, say, for 2-NEG-CNF. But this simply comes from the fact that the negation of a 2-POS-DNF formula is a 2-NEG-CNF formula, and from the fact that, for a Boolean classifier M we have that

$$\mathsf{SHAP}(M, \mathbf{e}, x) \quad = \quad -\mathsf{SHAP}(\neg M, \mathbf{e}, x). \tag{18}$$

This last property can be easily shown by considering that $\phi(M, \mathbf{e}, S) + \phi(\neg M, \mathbf{e}, S) = 1$, for every Boolean classifier M over a set of features X and every subset S of X. This concludes the proof of Corollary 10.

7. On Comparing the Values of the SHAP-Score

We now turn our attention to the problem of comparing the SHAP-score of the features of an entity. The reason we are interested in this problem is that it could be the case that computing the SHAP-score exactly or approximately is intractable (as we have shown in the last two sections), but yet that we are able to compare the relevance of features. In this section we will again use the uniform distribution and drop the subscripts $\mathcal{U}$. We now define the problem that we consider. Given a class $\mathcal{C}$ of Boolean classifiers, the input of the problem $\mathsf{CompSHAP}(\mathcal{C})$ is a Boolean classifier $M \in \mathcal{C}$ over a set of features X, an entity $\mathbf{e}$ over X and two features $x, y \in X$, and the question to answer is whether $\mathsf{SHAP}(M, \mathbf{e}, x) > \mathsf{SHAP}(M, \mathbf{e}, y)$.[8]

We prove that this problem is unlikely to be tractable, even for very restricted Boolean classifiers. A formula φ is said to be in 3-POS-DNF if φ is in POS-DNF and every disjunct in φ contains at most three variables. Then we have that:

Theorem 16 *Let $\mathcal{C}_{\text{3-POS-DNF}}$ be the class of Boolean classifiers given as 3-POS-DNF formulas. If $\mathsf{CompSHAP}(\mathcal{C}_{\text{3-POS-DNF}}) \in$ BPP, then* NP = RP.

Moreover, as in the previous section, Lemma 15 and Equation (18) directly imply that this result extends to the closure of 3-POS-DNF formulas by duals and negations:

Corollary 17 *Let $\mathcal{F} \in \{$3-POS-DNF, 3-NEG-DNF, 3-POS-CNF, 3-NEG-CNF$\}$ and $\mathcal{C}_{\mathcal{F}}$ be the class of Boolean classifiers given as formulas in $\mathcal{F}$. If $\mathsf{CompSHAP}(\mathcal{C}_{\mathcal{F}}) \in$ BPP, then* NP = RP.

Hence, all we have to do is to prove Theorem 16. The proof is again technical and divided in two parts, highlighted in bold in the rest of this section.

8. Note that $>$ can be replaced by any of $\geq, <$, or $\leq$, as this does not change the problem.

An intermediate problem: $\mathsf{DistCompSHAP}(\mathcal{C})$. As a first step, we consider a variation of the problem $\mathsf{CompSHAP}$. Let $\mathcal{C}$ be a class of Boolean classifiers. The input of the problem $\mathsf{DistCompSHAP}(\mathcal{C})$ is a pair of Boolean classifiers $M, M' \in \mathcal{C}$ over the same set of features X and the question to answer is whether

$$\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}, x) \quad > \quad \mathsf{SHAP}(M' \wedge \neg y, \mathbf{e}', y).$$

where x, y are two features not occurring in X, $\mathbf{e}$ is the entity over $X \cup \{x\}$ such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(z) = 1$ for every $z \in X$, and $\mathbf{e}'$ is the entity over $X \cup \{y\}$ such that $\mathbf{e}'(y) = 0$ and $\mathbf{e}'(z) = 1$ for every $z \in X$. We claim that $\mathsf{DistCompSHAP}$ over 3-NEG-CNF formulas can be reduced in polynomial time to $\mathsf{CompSHAP}$ over 2-POS-DNF formulas:

Lemma 18 *Let $\mathcal{C}_{\text{2-NEG-CNF}}$ be the class of Boolean classifiers given as 2-NEG-CNF formulas. There exists a polynomial-time many-one reduction from $\mathsf{DistCompSHAP}(\mathcal{C}_{\text{2-NEG-CNF}})$ to $\mathsf{CompSHAP}(\mathcal{C}_{\text{3-POS-DNF}})$.*

This in particular implies that if $\mathsf{CompSHAP}(\mathcal{C}_{\text{3-POS-DNF}})$ is in BPP then so is $\mathsf{DistCompSHAP}(\mathcal{C}_{\text{2-NEG-CNF}})$. We need three lemmas to prove Lemma 18.

Lemma 19 *Given a Boolean classifier M over a set of features X, an entity $\mathbf{e}$ over X and features $x, y \in X$, we have that:*

$$\mathsf{SHAP}(M, \mathbf{e}, x) - \mathsf{SHAP}(M, \mathbf{e}, y) \; = \; \sum_{S \subseteq X \setminus \{x,y\}} \frac{|S|!\,(|X| - |S| - 2)!}{(|X| - 1)!} \big(\phi(M, \mathbf{e}, S \cup \{x\}) - \phi(M, \mathbf{e}, S \cup \{y\})\big).$$

Proof We prove this in Appendix C.2. ■

Lemma 20 *Let X be a set of features, $n = |X|$, $x \in X$, M be a Boolean classifier over $X \setminus \{x\}$ and $\mathbf{e}$ be the entity over X such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(y) = 1$ for every $y \in X \setminus \{x\}$. Then we have that:*

$$\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}, x) \quad = \quad \sum_{k=0}^{n-1} \frac{k!(n-k-1)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq X \setminus \{x\} \,:\, |S| = k} \#\mathsf{SAT}(M, S).$$

Proof This can be established in the same way as Lemma 11 is proved. ■

Lemma 21 *Let X be a set of features, $n = |X|$, $x, y \in X$, M, M' be Boolean classifiers over $X \setminus \{x, y\}$ and $\mathbf{e}$ be the entity over X such that $\mathbf{e}(x) = 0$, $\mathbf{e}(y) = 0$ and $\mathbf{e}(z) = 1$ for every $z \in X \setminus \{x, y\}$. Then we have that:*

$$\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}_{|X \setminus \{y\}}, x) - \mathsf{SHAP}(M' \wedge \neg y, \mathbf{e}_{|X \setminus \{x\}}, y) \; = \; \mathsf{SHAP}((\neg M' \wedge x) \vee (\neg M \wedge y), \mathbf{e}, y) - \mathsf{SHAP}((\neg M' \wedge x) \vee (\neg M \wedge y), \mathbf{e}, x).$$

Proof Let $S \subseteq X \setminus \{x, y\}$. We have that:

$$\begin{aligned}
\phi((\neg M' \wedge x) \vee (\neg M \wedge y), \mathbf{e}, S \cup \{x\}) \; = \; & \\
& \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\})} \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \cdot ((\neg M' \wedge x) \vee (\neg M \wedge y))(\mathbf{e}') \; = \\
& \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\})} (\neg M \wedge y)(\mathbf{e}') \; = \\
& \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \Bigg(\sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\}) \,:\, \mathbf{e}'(y) = 1} (\neg M \wedge y)(\mathbf{e}') + \\
& \qquad \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\}) \,:\, \mathbf{e}'(y) = 0} (\neg M \wedge y)(\mathbf{e}') \Bigg) \; = \\
& \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}, S \cup \{x\}) \,:\, \mathbf{e}'(y) = 1} (\neg M)(\mathbf{e}') \; = \\
& \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \sum_{\mathbf{e}' \in \mathsf{cw}(\mathbf{e}_{|X \setminus \{x,y\}}, S)} (\neg M)(\mathbf{e}') \; = \\
& \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \cdot \#\mathsf{SAT}(\neg M, S).
\end{aligned}$$

Similarly we have that:

$$\phi((\neg M' \wedge x) \vee (\neg M \wedge y), \mathbf{e}, S \cup \{y\}) \;\; = \;\; \frac{1}{2^{|X \setminus (S \cup \{y\})|}} \cdot \#\mathsf{SAT}(\neg M', S).$$

Hence, we have from Lemma 19 (applied to the classifier $(\neg M' \wedge x) \vee (\neg M \wedge y)$) that:

$$\begin{aligned}
& \mathsf{SHAP}((\neg M' \wedge x) \vee (\neg M \wedge y), \mathbf{e}, y) - \mathsf{SHAP}((\neg M' \wedge x) \vee (\neg M \wedge y), \mathbf{e}, x) \; = \\
& \quad \sum_{S \subseteq X \setminus \{x,y\}} \frac{|S|!(|X| - |S| - 2)!}{(|X| - 1)!} \cdot \Bigg(\frac{1}{2^{|X \setminus (S \cup \{y\})|}} \cdot \#\mathsf{SAT}(\neg M', S) - \\
& \qquad \frac{1}{2^{|X \setminus (S \cup \{x\})|}} \cdot \#\mathsf{SAT}(\neg M, S) \Bigg) \; = \\
& \Bigg(\sum_{k=0}^{n-2} \frac{k!(n-k-2)!}{(n-1)!} \cdot \frac{1}{2^{n-k-1}} \sum_{S \subseteq X \setminus \{x,y\} \,:\, |S| = k} \#\mathsf{SAT}(\neg M', S) \Bigg) - \\
& \qquad \Bigg(\sum_{k=0}^{n-2} \frac{k!(n-k-2)!}{(n-1)!} \cdot \frac{1}{2^{n-k-1}} \sum_{S \subseteq X \setminus \{x,y\} \,:\, |S| = k} \#\mathsf{SAT}(\neg M, S) \Bigg) \; = \\
& \Bigg(\sum_{k=0}^{n-2} \frac{k!(n-k-2)!}{(n-1)!} \cdot \frac{1}{2^{n-k-1}} \sum_{S \subseteq X \setminus \{x,y\} \,:\, |S| = k} \big(2^{n-2-k} - \#\mathsf{SAT}(M', S)\big) \Bigg) - \\
& \qquad \Bigg(\sum_{k=0}^{n-2} \frac{k!(n-k-2)!}{(n-1)!} \cdot \frac{1}{2^{n-k-1}} \sum_{S \subseteq X \setminus \{x,y\} \,:\, |S| = k} \big(2^{n-2-k} - \#\mathsf{SAT}(M, S)\big) \Bigg) \; =
\end{aligned}$$

$$\left(\sum_{k=0}^{n-2} \frac{k!(n-k-2)!}{(n-1)!} \cdot \frac{1}{2^{n-k-1}} \sum_{S \subseteq X \setminus \{x,y\} \,:\, |S|=k} \#\mathsf{SAT}(M,S)\right) -$$
$$\left(\sum_{k=0}^{n-2} \frac{k!(n-k-2)!}{(n-1)!} \cdot \frac{1}{2^{n-k-1}} \sum_{S \subseteq X \setminus \{x,y\} \,:\, |S|=k} \#\mathsf{SAT}(M',S))\right)$$

We can now use Lemma 20 to conclude the proof.

■

Proof of Lemma 18 Let M, M' be Boolean classifiers in $\mathcal{C}_{\text{2-NEG-CNF}}$ over a set of variables X, which are inputs of DistCompSHAP. Remember that we want to decide whether

$$\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}, x) \;\; > \;\; \mathsf{SHAP}(M' \wedge \neg y, \mathbf{e}', y),$$

where x, y are two features not occurring in X, $\mathbf{e}$ is the entity over $X \cup \{x\}$ such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(z) = 1$ for every $z \in X$, and $\mathbf{e}'$ is the entity over $X \cup \{y\}$ such that $\mathbf{e}'(y) = 0$ and $\mathbf{e}'(z) = 1$ for every $z \in X$. Observe that $\neg M$ is a formula in 2-POS-DNF, and that $\neg M \wedge y$ is a formula in 3-POS-DNF (obtained by adding y to every term of $\neg M$). Similarly, $\neg M' \wedge x$ is a formula in 3-POS-DNF. Therefore, $(\neg M' \wedge x) \vee (\neg M \wedge y)$ is also a formula in 3-POS-DNF, and then Lemma 21 can be used to establish the polynomial-time many-one reduction. ■

This concludes the first part of the proof. Next, we show that DistCompSHAP is unlikely to be in BPP for the class of Boolean classifiers given as 2-NEG-CNF formulas.

Intractability of DistCompSHAP over 2-NEG-CNF. We now prove that if $\mathsf{DistCompSHAP}(\mathcal{C}_{\text{2-NEG-CNF}}) \in \text{BPP}$, then NP = RP. Notice that this implies that Theorem 16 holds, given that in the previous section we prove that if $\mathsf{CompSHAP}(\mathcal{C}_{\text{3-POS-DNF}}) \in \text{BPP}$, then $\mathsf{DistCompSHAP}(\mathcal{C}_{\text{2-NEG-CNF}}) \in \text{BPP}$.

In what follows, we consider the same class of graphs as in the proof of Theorem 9, that is, the class of all undirected and loop-free graphs $G = (N, E)$ containing at least two isolated nodes (see Condition (A)). Besides, recall the definition of the 2-POS-DNF formula $\theta(G)$ from Equation (16). We start our proof with the following lemma:

Lemma 22 *Let $G = (N, E)$ be a graph with $n = |N|$ nodes, x be a propositional variable not occurring in $\theta(G)$, and $\mathbf{e}$ be an entity such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(y) = 1$ for each variable y occurring in $\theta(G)$. Then we have that:*

$$\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x) \;\; = \;\; \frac{1}{2(n+1)} \sum_{k=0}^{n} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq N \,:\, |S|=k} \#\mathsf{CLIQUE}(G,S).$$

Proof The proof is similar to that of Lemma 12, but this time we use Lemma 20 instead of Lemma 11. ■

We will also need the fact that this quantity cannot be approximated:

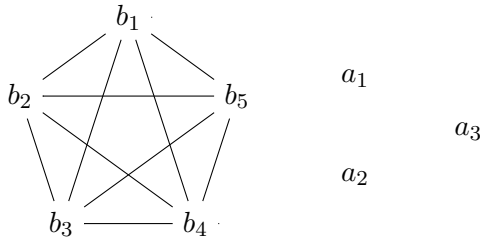


Figure 4: Illustration of the graph $G_{8,3}$.

Lemma 23 *For every $\varepsilon \in (0,1)$, the following problem does not admit an ε-PRA, unless* NP = RP*. Given as input a graph G, compute $\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)$, where x is a fresh variable not occurring in $\theta(G)$ and $\mathbf{e}$ is an entity such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(y) = 1$ for each variable y occurring in $\theta(G)$.*

Proof From the proof of Theorem 9, we get that there is no ε-PRA for the following problem, unless NP = RP. Given as input a graph G, compute the quantity:

$$\frac{1}{2(n+1)} \sum_{k=0}^{n} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq N \,:\, |S|=k} \#\mathsf{CLIQUE}(G, S).$$

Hence, Lemma 22 allows us to conclude that Lemma 23 holds. ■

The idea of our proof is then the following. We will prove that if $\mathsf{DistCompSHAP}(\mathcal{C}_{\text{2-NEG-CNF}})$ is in BPP, then there exists a $\frac{9}{10}$-PRA for the problem in Lemma 23, hence deducing that NP = RP. To this end, we introduce a class of graphs and calculate the values of $\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)$ over them. Given $n, t \in \mathbb{N}$ with $t \leq n$, we define the graph $G_{n,t} = (N_{n,t}, E_{n,t})$ with $N_{n,t} = \{a_1, \ldots, a_t\} \cup \{b_1, \ldots, b_{n-t}\}$ and $E_{n,t} = \{(b_i, b_j) \mid i, j \in \{1, \ldots n - t\} \text{ and } i \neq j\}$. In other words, $G_{n,t}$ is the disjoint union of a clique with $n-t$ nodes and of a graph consisting of t isolated nodes. For instance, the graph $G_{8,3}$ is illustrated in Figure 4. We calculate the values $\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x)$ in the next lemma.

Lemma 24 *Let $n, t \in \mathbb{N}$ such that $t \leq n$, x be a variable not occurring in $\theta(G_{n,t})$ and $\mathbf{e}$ an entity such $\mathbf{e}(x) = 0$ and $\mathbf{e}(y) = 1$ for every variable y occurring in $\theta(G_{n,t})$. Then:*

$$\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x) \quad = \quad \frac{t}{n(n+1)2^n} + \frac{1}{(t+1)2^{t+1}}.$$

Proof By Lemma 22 we have

$$\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x) \quad = \quad \frac{1}{2(n+1)} \sum_{k=0}^{n} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq N_{n,t} \,:\, |S|=k} \#\mathsf{CLIQUE}(G, S).$$

Let A be the set of the t isolated nodes of $G_{n,t}$ and B be the set of the $n-t$ nodes of $G_{n,t}$ that form a clique. In the above expression, it is clear that if S intersects both A and B then $\#\mathsf{CLIQUE}(G, S)$ is empty (because S is already not a clique). Moreover, if S is included

in A and is not empty, then $\#\mathsf{CLIQUE}(G, S)$ is 1 if $|S| = 1$ and is empty otherwise. Finally, if S is included in B then we have $\#\mathsf{CLIQUE}(G, S) = 2^{n-t-|S|}$. Thus, we obtain

$$\begin{aligned}
\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x) &= \frac{1}{2(n+1)} \frac{1!(n-1)!}{n!} \cdot \frac{1}{2^{n-1}} \sum_{S \subseteq A \,:\, |S|=1} 1 \\
&\quad + \frac{1}{2(n+1)} \sum_{k=0}^{n-t} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \sum_{S \subseteq B \,:\, |S|=k} 2^{n-t-k} \\
&= \frac{t}{n(n+1)2^n} \\
&\quad + \frac{1}{2(n+1)} \sum_{k=0}^{n-t} \frac{k!(n-k)!}{n!} \cdot \frac{1}{2^{n-k}} \binom{n-t}{k} 2^{n-t-k} \\
&= \frac{t}{n(n+1)2^n} + \frac{1}{2(n+1)2^t} \sum_{k=0}^{n-t} \frac{(n-t)!(n-k)!}{n!(n-t-k)!}
\end{aligned}$$

We prove in Appendix C.3 that the sum on the right is equal to $\frac{n+1}{t+1}$, so we get

$$\begin{aligned}
\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x) &= \frac{t}{n(n+1)2^n} + \frac{1}{2(n+1)2^t} \cdot \frac{n+1}{t+1} \\
&= \frac{t}{n(n+1)2^n} + \frac{1}{(t+1)2^{t+1}},
\end{aligned}$$

which was to be shown. ■

As a final ingredient, we will also need the following simple observation:

Lemma 25 *Let $G = (N, E)$ be a graph with $n = |N|$, x be a variable not occurring in $\theta(G)$ and $\mathbf{e}$ an entity such $\mathbf{e}(x) = 0$ and $\mathbf{e}(y) = 1$ for every variable y occurring in $\theta(G)$. Assuming, without loss of generality, that $G_{n,t}$ has the same nodes as G, for each $t \in \{0, \ldots, n\}$, we have that:*

$$\mathsf{SHAP}(\neg\theta(G_{n,n}) \wedge \neg x, \mathbf{e}, x) \;\leq\; \mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x) \;\leq\; \mathsf{SHAP}(\neg\theta(G_{n,2}) \wedge \neg x, \mathbf{e}, x).$$

Proof By looking at Lemma 22, we see that $\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)$ can only increase when edges are added to the graph G. Therefore, this quantity is lower bounded by the quantity for the graph with n isolated nodes (that is, $G_{n,n}$) and upper bounded by the quantity for the graph $G_{n,2}$, given that we only consider graphs with at least two isolated nodes. ■

We have the necessary ingredients to finish the proof of Theorem 16. Assume that there exists a BPP algorithm $\mathcal{A}$ for the problem $\mathsf{DistCompSHAP}(\mathcal{C}_{\text{2-NEG-CNF}})$. Then the input of $\mathcal{A}$ is a pair of Boolean classifiers M, M' over a set of features X given as 2-NEG-CNF formulas, and the task is to verify whether $\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}, x) > \mathsf{SHAP}(M' \wedge \neg y, \mathbf{e}', y)$,

where x, y are two features not occurring in X, $\mathbf{e}$ is an entity over $X \cup \{x\}$ such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(z) = 1$ for every $z \in X$, and $\mathbf{e}'$ is an entity over $X \cup \{y\}$ such that $\mathbf{e}'(y) = 0$ and $\mathbf{e}'(z) = 1$ for every $z \in X$. Using the amplification lemma of BPP (Goldreich, 2008, p. 231), we can ensure that $\mathcal{A}$ satisfies the following conditions:

- if $\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}, x) > \mathsf{SHAP}(M' \wedge \neg y, \mathbf{e}', y)$, then

$$\Pr(\mathcal{A}(M, M') \text{ outputs } \mathbf{yes}) \quad \geq \quad \left(1 - \frac{1}{4(\|M\| + \|M'\|)}\right).$$

- If $\mathsf{SHAP}(M \wedge \neg x, \mathbf{e}, x) \leq \mathsf{SHAP}(M' \wedge \neg y, \mathbf{e}', y)$, then

$$\Pr(\mathcal{A}(M, M') \text{ outputs } \mathbf{no}) \quad \geq \quad \left(1 - \frac{1}{4(\|M\| + \|M'\|)}\right).$$

Moreover, $\mathcal{A}(M, M')$ works in time $O(p(\|M\| + \|M'\|))$, where $p(\cdot)$ is a fixed polynomial, and $\|M\|$, $\|M'\|$ are the sizes of M and M' represented as input strings, respectively. By using BPP algorithm $\mathcal{A}$, we will define an algorithm $\mathcal{B}$ for approximating, given a graph $G = (N, E)$ containing at least 2 isolated nodes, the value $\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)$, where x is a feature not occurring in N and $\mathbf{e}$ is the entity over $N \cup \{x\}$ such that $\mathbf{e}(x) = 0$ and $\mathbf{e}(z) = 1$ for every $z \in N$. More precisely, we will show that $\mathcal{B}$ is a $\frac{9}{10}$-PRA for this problem. As mentioned above, this concludes the proof of Theorem 16 by Lemma 23 and Lemma 18. We now define $\mathcal{B}$.

Given a graph $G = (N, E)$ containing at least 2 isolated nodes, and assuming that $n = |N|$, algorithm $\mathcal{B}(G)$ performs the following steps:

1. If $n < 4$ then compute the exact value of $\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)$ by using an exhaustive approach.

2. For $t = 2$ to n:

   (a) if $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t})) = \mathbf{yes}$, then return

$$\frac{1}{2}\left(\mathsf{SHAP}(\neg\theta(G_{n,t-1}) \wedge \neg x, \mathbf{e}, x) + \mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x)\right)$$

3. Return $\mathsf{SHAP}(\neg\theta(G_{n,n}) \wedge \neg x, \mathbf{e}, x)$.

Algorithm $\mathcal{B}$ works in polynomial time since each graph $G_{n,t}$ can be constructed in polynomial time in n, 2-NEG-CNF formulas $\neg\theta(G)$ and $\neg\theta(G_{n,t})$ can be constructed in polynomial time from graphs G and $G_{n,t}$, algorithm $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t}))$ works in time $O(p(\|\neg\theta(G)\| + \|\neg\theta(G_{n,t})\|))$, and the value returned in either Step 2a or Step 3 can be computed in polynomial time in n by Lemma 24.

As the final step of this proof, we need to show that:

$$\Pr\left(\left|\mathcal{B}(G) - \mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)\right| \leq \frac{9}{10} \cdot \left|\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)\right|\right) \;\geq\; \frac{3}{4}. \qquad (19)$$

We can assume without loss of generality that $n \geq 4$, since otherwise the algorithm has computed the exact value. Assume initially that every call $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t}))$ in algorithm $\mathcal{B}$ returns the correct result (we will come back to this assumption later). Because the values $\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x)$ are strictly decreasing with t (this is clear by looking at the expression in Lemma 22), by Lemma 25 we have that either $\mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x) = \mathsf{SHAP}(\neg\theta(G_{n,n}) \wedge \neg x, \mathbf{e}, x)$ or there exists $t \in \{3, \ldots, n\}$ such that $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t})) = \mathbf{yes}$ and $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t-1})) = \mathbf{no}$ (notice that $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,2})) = \mathbf{no}$ since G contains at least 2 isolated nodes). In the former case, the third step of our algorithm ensures that we return the correct value, so let us focus on the latter case. For some $t \in \{3, \ldots, n\}$, we have that:

$$\mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x) \;<\; \mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x) \;\leq\; \mathsf{SHAP}(\neg\theta(G_{n,t-1}) \wedge \neg x, \mathbf{e}, x). \quad (20)$$

Therefore, by considering that the returned value in Step 2a of invocation $\mathcal{B}(G)$ is the average of the left and right terms of the above interval, we obtain that:

$$\Big|\mathcal{B}(G) - \mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)\Big| \;\leq\; \frac{1}{2}\Big(\mathsf{SHAP}(\neg\theta(G_{n,t-1}) \wedge \neg x, \mathbf{e}, x) - \mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x)\Big).$$

Now, we have by Lemma 24 that:

$$\begin{aligned}
\frac{1}{2}\Big(\mathsf{SHAP}(\neg\theta(G_{n,t-1}) \wedge \neg x, \mathbf{e}, x) - \mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x)\Big) \;&=\; \\
\frac{t-1}{n(n+1)2^{n+1}} + \frac{1}{t2^{t+1}} - \frac{t}{n(n+1)2^{n+1}} - \frac{1}{(t+1)2^{t+2}} \;&=\; \\
\frac{1}{t2^{t+1}} - \frac{1}{(t+1)2^{t+2}} - \frac{1}{n(n+1)2^{n+1}} \;&\leq\; \\
\frac{1}{t2^{t+1}} - \frac{1}{(t+1)2^{t+2}} \;&=\; \\
\frac{t+2}{2t} \cdot \frac{1}{(t+1)2^{t+1}} &
\end{aligned}$$

But notice that $\frac{t+2}{2t} = \frac{1}{2} + \frac{1}{t} \leq \frac{9}{10}$ since $t \geq 3$. Therefore we have

$$\begin{aligned}
\Big|\mathcal{B}(G) - \mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x)\Big| \;&\leq\; \frac{9}{10} \cdot \frac{1}{(t+1)2^{t+1}} \\
&\leq\; \frac{9}{10} \cdot \left(\frac{t}{n(n+1)2^{n}} + \frac{1}{(t+1)2^{t+1}}\right) \\
&=\; \frac{9}{10} \cdot \mathsf{SHAP}(\neg\theta(G_{n,t}) \wedge \neg x, \mathbf{e}, x) \\
&\leq\; \frac{9}{10} \cdot \mathsf{SHAP}(\neg\theta(G) \wedge \neg x, \mathbf{e}, x), \qquad (21)
\end{aligned}$$

where the last inequality comes from Equation (20).

To finish with the proof, we need to remove the assumption that every call $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t}))$ in algorithm $\mathcal{B}$ returns the correct result, and instead compute a lower bound on the probability that this assumption is true. More precisely, we need to show that the probability that every call $\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t}))$ in algorithm $\mathcal{B}$ returns the correct result is at least $\frac{3}{4}$, as this lower bound together with (21) imply that (19) holds (given that $\mathsf{SHAP}(\neg\theta(G)\wedge\neg x, \mathbf{e}, x) = |\mathsf{SHAP}(\neg\theta(G)\wedge\neg x, \mathbf{e}, x)|$ by Lemma 25). For every $t \in \{2, \ldots, n\}$, we have that:

$$\begin{aligned}\Pr(\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t})) \text{ outputs the correct result}) &\geq \left(1 - \frac{1}{4(\|M\| + \|M'\|)}\right) \\ &\geq \left(1 - \frac{1}{4n}\right)\end{aligned}$$

since $\|M\| + \|M'\| \geq n$. Hence,

$$\begin{aligned}\Pr\left(\bigwedge_{t=2}^{n} \left(\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t})) \text{ outputs the correct result}\right)\right) &\geq \left(1 - \frac{1}{4n}\right)^{n-1} \\ &\geq \left(1 - \frac{1}{4n}\right)^{n}\end{aligned}$$

But it is well known that the function $f : x \mapsto \left(1 - \frac{1}{4x}\right)^x$ is strictly increasing for $x \geq 1$, and that $f(4) > \frac{3}{4}$. Hence, since we assumed $n \geq 4$, we obtain

$$\Pr\left(\bigwedge_{t=2}^{n} \left(\mathcal{A}(\neg\theta(G), \neg\theta(G_{n,t})) \text{ outputs the correct result}\right)\right) \geq \frac{3}{4}.$$

This concludes the proof of Theorem 16.

8. Final Remarks

Our algorithm for computing the SHAP-score could be used in practical scenarios. Indeed, some classes of classifiers can be compiled into tractable Boolean circuits. This is the case, for instance, of Bayesian Classifiers (Shih et al., 2018b), Binary Neural Networks (Shi et al., 2020), and Random Forests (Choi et al., 2020). The idea is to start with a Boolean classifier M given in a formalism that is hard to interpret – for instance a binary neural network – and to compute a tractable Boolean circuit M' that is equivalent to M (this computation can be expensive). One can then use M' and the nice properties of tractable Boolean circuits to explain the decisions of the model. Hence, this makes it possible to apply the results in this paper on the SHAP-score to those classes of classifiers.

A Boolean circuit representing (or compiled from) another opaque classifier, as those just mentioned, is also more interpretable (Rudin, 2019), adding a useful property to that of tractability. Still, it would be interesting to compare the explanations obtained from the compiled circuit with those that can be obtain directly from, say, the original neural network. After all, the SHAP-score can be applied to both. First and recent experiments of this kind are reported in Bertossi and Leon (2023). We should mention, however, that there are some recent methods to explain the results from a neural network that do not rely only

on its input/output relation (c.f. (Samek et al., 2021) for a recent and thorough review). Quite recent approaches to the identification and modeling of causal structures in deep learning should open the door for new methodologies for interpretation and explanation (Ahmed et al., 2020; Schölkopf et al., 2021).

To a large extent, explanations in ML are based on different forms of counterfactual- or causality-based approaches; and the concept of explanation as such is left rather implicit. However, explanations have been treated in more explicit terms in many disciplines, and, in particular, in AI, under *consistency-based* and *abductive* diagnosis, the main two forms of *model-based diagnosis* (Struss, 2008). These are explanations in knowledge representation with open models. Recent progress has been made in applying model-based diagnosis in Explainable ML. C.f. Bertossi (2023) and Marques-Silva (2022) for technical details and references. However, a deeper investigation of this connection in the context of explanations for classification is open.

We conclude this discussion by mentioning a few research directions that our work opens. First, it would be interesting to extend our tractability results of Sections 3 and 4 to a setting that could allow for *continuous* variables, instead of the discrete variables that we have considered so far. A starting point for such an investigation could for instance be the framework of *probabilistic circuits* proposed by Van den Broeck et al. (2019): these are data structures that can be used to represent probability distributions (possibly with continuous variables) so as to ensure the tractability of certain tasks, such as expectation computation or probabilistic inference. Since these circuits are very similar in nature to the deterministic and decomposable circuit classes that we consider here, it seems that they could also be used for the SHAP-score. Another natural direction would be to study the complexity of approximately computing the SHAP-score under so-called *empirical distributions*, as is done in Van den Broeck et al. (2021, 2022) for the case of exact computation where they show that it is intractable in general. For instance, is it the case that there is still no FPRAS for approximating the SHAP-score of DNF formulas (or monotone DNFs) under these distributions? Last, we leave open the question of finding a natural class of Boolean classifiers for which we could efficiently approximate the SHAP-score (via an FPRAS), or efficiently compare the SHAP-score of different features, whereas computing this score exactly would be intractable. As we mentioned in Section 6, given that exact model counting for DNF formulas is intractable but has an FPRAS, and given the strong connections between model counting and the SHAP-score established in Section 5 or by Van den Broeck et al. (2021, 2022), DNF formulas were a natural candidate for this. But, quite surprisingly, our non-approximability results indicate that this is not the case.

Acknowledgments

Part of this work has been funded by ANID - Millennium Science Initiative Program - Code ICN17002. M. Arenas has been funded by Fondecyt grant 1191337. P. Barceló has been funded by Fondecyt Grant 1200967, and also by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. L. Bertossi is a Senior Researcher at the IMFD, Chile, and a Professor Emeritus at Carleton University, Ottawa, Canada.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wüthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020. URL https://arxiv.org/abs/2010.04296.

Antoine Amarilli, Florent Capelli, Mikaël Monet, and Pierre Senellart. Connecting knowledge compilation classes and width parameters. *Theory Comput. Syst.*, 64(5):861–914, 2020. URL https://arxiv.org/abs/1811.02944.

Marcelo Arenas, Pablo Barceló, Leopoldo Bertossi, and Mikaël Monet. The tractability of SHAP-score-based explanations over deterministic and decomposable boolean circuits. In *Proceedings of AAAI*, 2021. URL https://arxiv.org/abs/2007.14045.

Sanjeev Arora and Boaz Barak. *Computational Complexity - A Modern Approach*. Cambridge University Press, 2009. URL https://theory.cs.princeton.edu/complexity/book.pdf.

Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the ACM (JACM)*, 45(1):70–122, 1998. URL https://www.cs.umd.edu/users/gasarch/TOPICS/pcp/AS.pdf.

Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM (JACM)*, 45(3):501–555, 1998. URL https://madhu.seas.harvard.edu/papers/1992/almss-conf.pdf.

Leopoldo Bertossi. Attribution-scores and causal counterfactuals as explanations in artificial intelligence. *CoRR*, abs/2303.02829, 2023. doi: 10.48550/arXiv.2303.02829. URL https://doi.org/10.48550/arXiv.2303.02829. To appear in Reasoning Web. Causality, Explanations and Declarative Knowledge, Springer LNCS 13759, 2023.

Leopoldo Bertossi and Jorge E. Leon. Opening up the neural network classifier for Shap score computation. *CoRR*, abs/2303.06516, 2023. doi: 10.48550/arXiv.2303.06516. URL https://doi.org/10.48550/arXiv.2303.06516.

Leopoldo Bertossi, Jordan Li, Maximilian Schleich, Dan Suciu, and Zografoula Vagena. Causality-based explanation of classification outcomes. In *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, pages 1–10, 2020. URL https://arxiv.org/abs/2003.06868.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Giulia Cesari, Encarnación Algaba, Stefano Moretti, and Juan A. Nepomuceno. An application of the Shapley value to the analysis of co-expression networks. *Applied Network Science*, 3(1):35, 2018. URL https://hal.archives-ouvertes.fr/hal-02103359.

Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615*, 2018. URL https://arxiv.org/abs/1811.12615.

Arthur Choi, Andy Shih, Anchal Goyanka, and Adnan Darwiche. On symbolically encoding the behavior of random forests. *arXiv preprint arXiv:2007.01493*, 2020. URL https://arxiv.org/abs/2007.01493.

Ian Covert and Su-In Lee. Improving KernelSHAP: Practical Shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2021. URL https://arxiv.org/abs/2012.01536.

Adnan Darwiche. On the tractable counting of theory models and its application to truth maintenance and belief revision. *Journal of Applied Non-Classical Logics*, 11(1-2):11–34, 2001. URL https://arxiv.org/abs/cs/0003044.

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *ECAI*, pages 712–720. IOS Press, 2020. URL https://arxiv.org/abs/2002.09284.

Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17:229–264, 2002. URL https://arxiv.org/abs/1106.1819.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016. URL https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf.

Xiaotie Deng and Christos H Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994.

Daniel Deutch, Nave Frost, Amir Gilad, and Oren Sheffer. Explanations for data repair through Shapley values. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 362–371, 2021. URL https://amirgilad.github.io/publication/cikm21/CIKM21.pdf.

Daniel Deutch, Nave Frost, Benny Kimelfeld, and Mikaël Monet. Computing the Shapley value of facts in query answering. In *Proceedings of the 2022 International Conference on Management of Data*, pages 1570–1583, 2022. URL https://arxiv.org/abs/2112.08874.

Ulrich Faigle and Walter Kern. The Shapley value for cooperative games under precedence constraints. *International Journal of Game Theory*, 21(3):249–266, 1992.

Uriel Feige, Shafi Goldwasser, László Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM (JACM)*, 43(2):268–292, 1996. URL http://ftp.cs.elte.hu/~lovasz/morepapers/fglss.pdf.

Gil Fidel, Ron Bitton, and Asaf Shabtai. When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. In *2020 International*

Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2020. URL https://arxiv.org/abs/1909.03418.

Oded Goldreich. *Computational complexity: a conceptual perspective*. Cambridge University Press, 2008. URL http://www.wisdom.weizmann.ac.il/~oded/cc-book.html.

Anthony Hunter and Sébastien Konieczny. On the measure of conflicts: Shapley inconsistency values. *Artificial Intelligence*, 174(14):1007–1026, 2010. URL http://www.cril.univ-artois.fr/~konieczny/papers/aij10a.pdf.

Richard M Karp, Michael Luby, and Neal Madras. Monte-carlo approximation algorithms for enumeration problems. *Journal of algorithms*, 10(3):429–448, 1989. URL https://www.math.cmu.edu/~af1p/Teaching/MCC17/Papers/KLM.pdf.

Ker-I Ko. Some observations on the probabilistic algorithms and NP-hard problems. *Information Processing Letters*, 14(1):39–43, 1982.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020. URL https://arxiv.org/abs/2002.11097.

Ester Livshits, Leopoldo Bertossi, Benny Kimelfeld, and Moshe Sebag. The Shapley value of tuples in query answering. In *ICDT*, volume 155 of *LIPIcs*, pages 20:1–20:19, 2020. URL https://arxiv.org/abs/1904.08679.

Ester Livshits, Leopoldo Bertossi, Benny Kimelfeld, and Moshe Sebag. The Shapley value of tuples in query answering. *Log. Methods Comput. Sci.*, 17(3), 2021. doi: 10.46298/lmcs-17(3:22)2021. URL https://doi.org/10.46298/lmcs-17(3:22)2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774, 2017. URL https://arxiv.org/abs/1705.07874.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. URL https://arxiv.org/abs/1905.04610.

João Marques-Silva. Logic-based explainability in machine learning. *CoRR*, abs/2211.00541, 2022. doi: 10.48550/arXiv.2211.00541. URL https://doi.org/10.48550/arXiv.2211.00541. To appear in Reasoning Web. Causality, Explanations and Declarative Knowledge, Springer LNCS 13759, 2023.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using Shapley values. In *CD-MAKE*, volume 12279 of *Lecture Notes in Computer Science*, pages 17–38. Springer, 2020. URL https://arxiv.org/abs/1909.08128.

Tomasz P Michalak, Karthik V Aadithya, Piotr L Szczepanski, Balaraman Ravindran, and Nicholas R Jennings. Efficient computation of the Shapley value for game-theoretic

network centrality. *Journal of Artificial Intelligence Research*, 46:607–650, 2013. URL https://arxiv.org/abs/1402.0567.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for Shapley value estimation. 2022. URL https://www.jmlr.org/papers/volume23/21-0439/21-0439.pdf.

C. Molnar. *Interpretable machine learning: A guide for making black box models explainable*. 2020. URL https://christophm.github.io/interpretable-ml-book/.

Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate Shapley values. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7992–7999. IEEE, 2021. URL https://arxiv.org/abs/2010.12082.

Robert Peharz, Steven Lang, Antonio Vergari, Karl Stelzner, Alejandro Molina, Martin Trapp, Guy Van den Broeck, Kristian Kersting, and Zoubin Ghahramani. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. *arXiv preprint arXiv:2004.06231*, 2020.

J Scott Provan and Michael O Ball. The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM Journal on Computing*, 12(4):777–788, 1983.

Shubham Rathi. Generating counterfactual and contrastive explanations using SHAP. *arXiv preprint arXiv:1906.09293*, 2019. URL http://arxiv.org/abs/1906.09293.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144. ACM, 2016. URL https://arxiv.org/abs/1602.04938.

Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988. URL http://www.library.fa.ru/files/Roth2.pdf.

Cynthia C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. URL https://arxiv.org/abs/1811.10154.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. URL https://doi.org/10.1109/JPROC.2021.3060483.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. URL https://arxiv.org/abs/2102.11107.

Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2 (28):307–317, 1953. URL http://www.library.fa.ru/files/Roth2.pdf#page=39.

Weijia Shi, Andy Shih, Adnan Darwiche, and Arthur Choi. On tractable representations of binary neural networks. In *KR*, pages 882–892, 2020. URL https://arxiv.org/abs/2004.02082.

Andy Shih, Arthur Choi, and Adnan Darwiche. Formal verification of Bayesian network classifiers. In *International Conference on Probabilistic Graphical Models*, pages 427–438, 2018a. URL http://proceedings.mlr.press/v72/shih18a.html.

Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining Bayesian network classifiers. In *IJCAI*, pages 5103–5111, 2018b. URL https://arxiv.org/abs/1805.03364.

Andy Shih, Adnan Darwiche, and Arthur Choi. Verifying binarized neural networks by Angluin-style learning. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 354–370. Springer, 2019a. URL http://web.cs.ucla.edu/~andyshih/assets/pdf/SDCsat19.pdf.

Andy Shih, Guy Van den Broeck, Paul Beame, and Antoine Amarilli. Smoothing structured decomposable circuits. In *NeurIPS*, pages 11412–11422, 2019b. URL https://arxiv.org/abs/1906.00311.

Alistair Sinclair. *Algorithms for random generation and counting - a Markov chain approach*. Progress in theoretical computer science. Birkhäuser, 1993. ISBN 978-0-8176-3658-6.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010. URL https://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf.

Peter Struss. Model-based problem solving. *Foundations of Artificial Intelligence*, 3:395–465, 2008. URL https://cse.sc.edu/~mgv/csce781sp13/presentations/struss-handbook-chapter.pdf.

Naoya Takeishi and Yoshinobu Kawahara. On anomaly interpretation via Shapley values. *arXiv preprint arXiv:2004.04464*, 2020. URL https://arxiv.org/abs/2004.04464.

Leslie G Valiant. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979. URL https://core.ac.uk/download/pdf/82500417.pdf.

Guy Van den Broeck, Nicola Di Mauro, and Antonio Vergari. Tractable probabilistic models: Representations, algorithms, learning, and applications, 2019. URL http://web.cs.ucla.edu/~guyvdb/slides/TPMTutorialUAI19.pdf. Tutorial at UAI 2019.

Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. On the tractability of SHAP explanations. In *Proceedings of AAAI*, 2021. URL https://arxiv.org/abs/2009.08634.

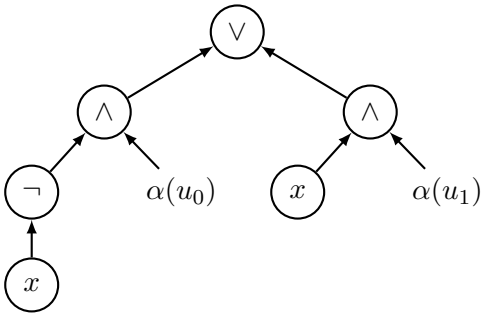
Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, 74:851–886, 2022. URL https://www.jair.org/index.php/jair/article/view/13283.

Appendix A. Encoding Binary Decision Trees and FBDDs into Deterministic and Decomposable Boolean Circuits

In this section we explain how FBDDs and binary decision trees can be encoded in linear time as deterministic and decomposable Boolean circuits. First we need to define these formalisms.

Binary Decision Diagrams A *binary decision diagram* (BDD) over a set of variables X is a rooted directed acyclic graph D such that: (i) each internal node is labeled with a variable from X, and has exactly two outgoing edges: one labeled 0, the other one labeled 1; and (ii) each leaf is labeled either 0 or 1. Such a BDD represents a Boolean classifier in the following way. Let $\mathbf{e}$ be an entity over X, and let $\pi_{\mathbf{e}} = u_1, \ldots, u_m$ be the unique path in D satisfying the following conditions: (a) u_1 is the root of D; (b) u_m is a leaf of D; and (c) for every $i \in \{1, \ldots, m-1\}$, if the label of u_i is $x \in X$, then the label of the edge (u_i, u_{i+1}) is equal to $\mathbf{e}(x)$. Then the *value of* $\mathbf{e}$ *in* D, denoted by $D(\mathbf{e})$, is defined as the label of the leaf u_m. Moreover, a binary decision diagram D is *free* (FBDD) if for every path from the root to a leaf, no two nodes on that path have the same label, and *a binary decision tree* is an FBDD whose underlying graph is a tree.

Encoding FBDDs and binary decision trees into deterministic and decomposable Boolean circuits (Folklore). Given an FBDD D over a set of variables X, we explain how D can be encoded as a deterministic and decomposable Boolean circuit C over X. Notice that the technique used in this example also apply to binary decision trees, as they are a particular case of FBDDs. The construction of C is done by traversing the structure of D in a bottom-up manner. In particular, for every node u of D, we construct a deterministic and decomposable circuit $\alpha(u)$ that is equivalent to the FBDD represented by the subgraph of D rooted at u. More precisely, for a leaf u of D that is labeled with $\ell \in \{0, 1\}$, we define $\alpha(u)$ to be the Boolean circuit consisting of only one constant gate with label ℓ. For an internal node u of D labeled with variable $x \in X$, let u_0 and u_1 be the nodes that we reach from u by following the 0- and 1-labeled edge, respectively. Then $\alpha(u)$ is the Boolean circuit depicted in the following figure:



It is clear that the circuit that we obtain is equivalent to the input FBDD. We now argue that this circuit is deterministic and decomposable. For the $\vee$-gate shown in the figure, if an entity $\mathbf{e}$ is accepted by the Boolean circuit in its left-hand size, then $\mathbf{e}(x) = 0$, while if an entity $\mathbf{e}$ is accepted by the Boolean circuit in its right-hand size, then $\mathbf{e}(x) = 1$. Hence, we have that this $\vee$-gate is deterministic, from which we conclude that $\alpha(u)$ is deterministic,

as $\alpha(u_0)$ and $\alpha(u_1)$ are also deterministic by construction. Moreover, the $\wedge$-gates shown in the figure are decomposable as variable x is mentioned neither in $\alpha(u_0)$ nor in $\alpha(u_1)$: this is because D is a *free* BDD. Thus, we conclude that $\alpha(u)$ is decomposable, as $\alpha(u_0)$ and $\alpha(u_1)$ are decomposable by construction. Finally, if u_{root} is the root of D, then by construction we have that $\alpha(u_{\mathsf{root}})$ is a deterministic and decomposable Boolean circuit equivalent to D. Note that this encoding can trivially be done in linear time. Thus, we often say, by abuse of terminology, that "FBDDs (or binary decision trees) are restricted kinds of deterministic and decomposable circuits".

Appendix B. Proof of a folklore fact

Fact 26 (Folklore) *Let f be a function that admits an ε-PRA for $\varepsilon \in (0,1)$. Then the problem of determining, given a string x, whether $f(x) = 0$ is in* BPP.

Proof Let $\mathcal{A}$ be an ε-PRA for f, that is, a randomized algorithm that takes as input a string x, and computes in polynomial time in the size of x a value $\mathcal{A}(x)$ such that $(\star)$ $\Pr\big(|f(x) - \mathcal{A}(x)| \leq \varepsilon\,|f(x)|\big) \geq \frac{3}{4}$. We claim that the following algorithm is a BPP algorithm for deciding if $f(x) = 0$: compute $\mathcal{A}(x)$, and if this is equal to zero then accept, otherwise reject. We will assume without loss of generality that $f(x) \geq 0$, as the case $f(x) \leq 0$ can be handled in the same way. Observe that $(\star)$ can be equivalently rewritten as $(\dagger)$ $\Pr\big((1-\varepsilon)f(x) \leq \mathcal{A}(x) \leq (1+\varepsilon)f(x)\big) \geq \frac{3}{4}$. But then:

- Assume first that $f(x) = 0$. Then by $(\dagger)$ we have that $\Pr\big(\mathcal{A}(x) = 0\big) \geq \frac{3}{4}$ as well, so that we accept with probability at least $\frac{3}{4}$.
- Assume now that $f(x) \neq 0$. Then $(\dagger)$ gives us $\Pr\big(\mathcal{A}(x) \geq (1-\varepsilon)f(x) > 0\big) \geq \frac{3}{4}$, so that we reject with probability at least $\frac{3}{4}$.

This concludes the proof. ■

Appendix C. Proofs of Intermediate Results

C.1 Proof of Lemma 14

We notice that the inequality holds if and only if

$$\frac{1+\varepsilon}{1-\varepsilon} \cdot (n^3+1) \cdot 2^n \quad < \quad 2^{mn^2 - 2\lfloor \frac{m}{3} \rfloor n^2}.$$

Moreover, given that $m \geq 1$, we have that:

$$\begin{aligned} 2^{mn^2 - 2\lfloor \frac{m}{3} \rfloor n^2} &\geq 2^{mn^2 - \frac{2m}{3} n^2} \\ &= 2^{\frac{m}{3} n^2} \\ &\geq 2^{\frac{n^2}{3}}. \end{aligned}$$

Therefore, we conclude that the inequality in the statement of the lemma holds from the fact that there exists z_0 such that:

$$\frac{1+\varepsilon}{1-\varepsilon} \cdot (z^3+1) \cdot 2^z \;<\; 2^{\frac{z^2}{3}} \qquad \text{for every } z \geq z_0.$$

This is indeed clear, as the dominating factor of the left term is $2^{O(z)}$ whereas that of the the right term is $2^{O(z^2)}$.

C.2 Proof of Lemma 19

We prove the more general claim that for any Boolean classifier M over features X, probability distribution $\mathcal{D}$ over $\mathsf{ent}(X)$, entity $\mathbf{e} \in \mathsf{ent}(X)$ and features $x, y \in X$, we have

$$\mathsf{SHAP}_{\mathcal{D}}(M,\mathbf{e},x) - \mathsf{SHAP}_{\mathcal{D}}(M,\mathbf{e},y) \;=\; \sum_{S \subseteq X \setminus \{x,y\}} \frac{|S|!\,(|X|-|S|-2)!}{(|X|-1)!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{y\})\big).$$

Indeed, we have:

$$\begin{aligned}\mathsf{SHAP}_{\mathcal{D}}(M,\mathbf{e},x) - \mathsf{SHAP}_{\mathcal{D}}(M,\mathbf{e},y) \;=\; & \sum_{S \subseteq X \setminus \{x\}} \frac{|S|!\,(|X|-|S|-1)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S)\big) \\ & - \sum_{S \subseteq X \setminus \{y\}} \frac{|S|!\,(|X|-|S|-1)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{y\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S)\big)\end{aligned}$$

We cut the first sum in half by considering (1) those S that are $\subseteq X \setminus \{x,y\}$ and (2) those $S \subseteq X$ such that with $y \in S$ and $x \notin S$; and do the same for the second sum. We end up with

$$\begin{aligned}\mathsf{SHAP}_{\mathcal{D}}(M,\mathbf{e},x) - \mathsf{SHAP}_{\mathcal{D}}(M,\mathbf{e},y) \;=\; & \sum_{S \subseteq X \setminus \{x,y\}} \frac{|S|!\,(|X|-|S|-1)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{y\})\big) \\ & + \sum_{S \subseteq X \setminus \{x,y\}} \frac{(|S|+1)!\,(|X|-|S|-2)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x,y\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{y\})\big) \\ & - \sum_{S \subseteq X \setminus \{x,y\}} \frac{(|S|+1)!\,(|X|-|S|-2)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x,y\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x\})\big) \\ = \; & \sum_{S \subseteq X \setminus \{x,y\}} \frac{|S|!\,(|X|-|S|-1)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{y\})\big) \\ & + \sum_{S \subseteq X \setminus \{x,y\}} \frac{(|S|+1)!\,(|X|-|S|-2)!}{|X|!}\big(\phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{x\}) - \phi_{\mathcal{D}}(M,\mathbf{e},S\cup\{y\})\big)\end{aligned}$$

$$= \sum_{S \subseteq X \setminus \{x,y\}} \frac{|S|!\,(|X| - |S| - 2)!}{(|X| - 1)!} \big(\phi_{\mathcal{D}}(M, \mathbf{e}, S \cup \{x\}) - \phi_{\mathcal{D}}(M, \mathbf{e}, S \cup \{y\})\big).$$

This concludes the proof.

C.3 The summation in Lemma 24

To ease the proof, we let $s = n - t$. Then we want to compute

$$\sum_{k=0}^{n-t} \frac{(n-t)!(n-k)!}{n!(n-t-k)!} \;=\; \sum_{k=0}^{s} \frac{s!(s+t-k)!}{(s+t)!(s-k)!}.$$

We have that:

$$\begin{aligned}
\sum_{k=0}^{s} \frac{s!(s+t-k)!}{(s+t)!(s-k)!} &= \frac{s!}{(s+t)!} \cdot \sum_{k=0}^{s} \frac{(s+t-k)!}{(s-k)!} \\
&= \frac{s!t!}{(s+t)!} \cdot \sum_{k=0}^{s} \frac{(s+t-k)!}{t!(s-k)!} \\
&= \frac{s!t!}{(s+t)!} \cdot \sum_{k=0}^{s} \binom{s+t-k}{s-k} \\
&= \frac{s!t!}{(s+t)!} \cdot \sum_{k=0}^{s} \binom{t+k}{k}. \qquad (22)
\end{aligned}$$

Next, we show that the following equality holds by induction on s:

$$\sum_{k=0}^{s} \binom{t+k}{k} \;=\; \binom{s+t+1}{s}. \qquad (23)$$

First, consider $s = 0$:

$$\sum_{k=0}^{s} \binom{t+k}{k} \;=\; \sum_{k=0}^{0} \binom{t+k}{k} \;=\; \binom{t+0}{0} \;=\; 1 \;=\; \binom{0+t+1}{0} \;=\; \binom{s+t+1}{s}.$$

Now, assuming that (23) holds, we have that:

$$\begin{aligned}
\sum_{k=0}^{s+1} \binom{t+k}{k} &= \sum_{k=0}^{s} \binom{t+k}{k} + \binom{t+s+1}{s+1} \\
&= \binom{s+t+1}{s} + \binom{t+s+1}{s+1} \qquad \text{by (23)} \\
&= \binom{s+t+2}{s+1} \qquad \text{by Pascal's rule} \\
&= \binom{s+1+t+1}{s+1}.
\end{aligned}$$

Hence, we conclude from (22) and 23 that:

$$\begin{aligned}\sum_{k=0}^{s} \frac{s!(s+t-k)!}{(s+t)!(s-k)!} &= \frac{s!t!}{(s+t)!} \cdot \binom{s+t+1}{s} \\ &= \frac{s!t!}{(s+t)!} \cdot \frac{(s+t+1)!}{s!(t+1)!} \\ &= \frac{s+t+1}{t+1},\end{aligned}$$

and this is $\frac{n+1}{t+1}$, as promised.