ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240586

The Distributional Uncertainty of the SHAP Score in Explainable Machine Learning

Santiago Cifuentes[a,d,*], **Leopoldo Bertossi**[b,1], **Nina Pardal**[c,a], **Sergio Abriola**[a,d], **Maria Vanina Martinez**[e] and **Miguel Romero**[f]

[a]Instituto de Ciencias de la Computación, UBA-CONICET, Argentina
[b]Universidad San Sebastián, FIAD, Santiago, Chile
[c]Department of Computer Science, University of Sheffield, UK
[d]Department of Computer Science, Universidad de Buenos Aires, Argentina
[e]Artificial Intelligence Research Institute (IIIA-CSIC), Spain
[f]Department of Computer Science, Universidad Católica de Chile, Chile
[f]CENIA Chile
ORCID (Santiago Cifuentes): https://orcid.org/0000-0002-6375-2045, ORCID (Leopoldo Bertossi): https://orcid.org/0000-0002-1144-3179, ORCID (Nina Pardal): https://orcid.org/0000-0002-5150-6947, ORCID (Sergio Abriola): https://orcid.org/0000-0002-1979-5443, ORCID (Maria Vanina Martinez): https://orcid.org/0000-0003-2819-4735, ORCID (Miguel Romero): https://orcid.org/0000-0002-2615-6455

Abstract. Attribution scores reflect how important the feature values in an input entity are for the output of a machine learning model. One of the most popular attribution scores is the SHAP score, which is an instantiation of the general Shapley value used in coalition game theory. The definition of this score relies on a probability distribution on the entity population. Since the exact distribution is generally unknown, it needs to be assigned subjectively or be estimated from data, which may lead to misleading feature scores. In this paper, we propose a principled framework for reasoning on SHAP scores under unknown entity population distributions. In our framework, we consider an uncertainty region that contains the potential distributions, and the SHAP score of a feature becomes a function defined over this region. We study the basic problems of finding maxima and minima of this function, which allows us to determine tight ranges for the SHAP scores of all features. In particular, we pinpoint the complexity of these problems, and other related ones, showing them to be intractable. Finally, we present experiments on a real-world dataset, showing that our framework may contribute to a more robust feature scoring.

1 Introduction

Proposing and investigating different forms of explaining and interpreting the outcomes from AI-based systems has become an effervescent area of research and applications [18], leading to the emergence of the area of Explainable Machine Learning. In particular, one wants to explain results obtained from ML-based classification systems. A widespread approach to achieve this consists in assigning numerical *attribution scores* to the feature values that, together, represent a given entity under classification, for which a given label has been obtained. The score of a feature value indicates how relevant it is for this output label.

One of the most popular attribution scores is the so-called *SHAP score* [16, 22], which is a particular form of the general *Shapley value* used in coalition game theory [23, 21]. SHAP, as every instantiation of the Shapley value, requires a *wealth function* shared by all the players in the coalition game. SHAP uses one that is an expected value based on a probability distribution on the entity population.[2] Since the exact distribution is generally unknown, it needs to be assigned subjectively or be estimated from data. This may lead to different kinds of errors, and in particular, to misleading feature scores.

In this work, we propose a principled framework for reasoning on SHAP scores under distributional uncertainty, that is, under an unknown distribution over the entity population. We focus on *binary classifiers*, i.e., classifiers that returns 1 (accept) or 0 (reject). We also assume that the inputs to these classifiers are *binary features*, i.e., features that can take values 0 or 1. Furthermore, we focus on product distributions. Their use for SHAP computation is common, and imposes feature independence [3, 27]. In practice, one frequently uses an *empirical product distribution* (of the empirical marginals), which may vary depending on the data set from which the sampling is performed [7]. We see our concentration on product distributions as an important first step towards the distributional analysis of SHAP.

Our approach allows us to reinterpret and analyze SHAP as a *function* defined on the uncertainty region. As it turns out, this function is always a polynomial on n variables (where n is the number of features), and hence we refer to it as the *SHAP polynomial*. We can then analyze the behavior of this polynomial to gain concrete insights on

* Corresponding Author. Email: scifuentes@dc.uba.ar
[1] Prof. Emeritus, Carleton Univ., Ottawa, Canada. bertossi@scs.carleton.ca. Senior Researcher IMFD, Chile.

[2] Since SHAP's inception, several variations have been proposed and investigated, but they all rely on some probability distribution. In this work we stick to the original formulation.

the importance of a feature.

Example 1. *Consider the classifier M given in Table 1. Let e be the null entity (first row), and assume a product distribution $\langle p_x, p_y, p_z\rangle$ over the feature space, e.g., $\mathbb{P}(x=1,y=0,z=1) = p_x(1-p_y)p_z$. The SHAP score for entity e and feature z depends on the probabilities p_x, p_y and p_z, and this relation can be expressed through the following function:*

$$\begin{aligned} Shap(M,e,z) &= Shap_{M,e,z}(p_x,p_y,p_z) \\ &= \frac{1}{6}p_z(-4p_xp_y+3p_x+3p_y). \end{aligned} \quad (1)$$

We call this function the SHAP polynomial for entity e and feature z (details are provided in Section 2). Observe that the term $(-4p_xp_y+3p_x+3p_y)$ is strictly positive whenever $p_x, p_y \neq 0$, and in those cases the SHAP score for z grows when p_z grows. This is intuitive: as p_z grows the probability that $e'(z)=1$ for a randomly chosen entity e' increases and this predisposes the classifier towards rejection (three out of four entities with $z=1$ are rejected by M). Therefore, the fact $e(z)=0$ becomes more informative, and consequently the SHAP score increases. Meanwhile, if $p_x = p_y = 0$ then the prediction is 1 with probability 1 independently of the value of z, and the SHAP score of z is 0.

x	y	z	M
0	0	0	**1**
0	0	1	**1**
0	1	0	**1**
1	0	0	**1**

Table 1: Classifier M, it labels the remaining entities with **0**.

There are different kinds of analysis one can carry out on the SHAP polynomial. As a first step, in this work we investigate the basic problem of finding maxima and minima of SHAP scores in the given region. This allows us to compute *SHAP intervals* for each feature: a range of all the values that the SHAP score can attain in the uncertainty region. We believe these tight ranges to be a valuable tool for reasoning about feature importance under distributional uncertainty. For instance, the length of the interval for a given feature provides information about the robustness of SHAP for that feature. Furthermore, changes of sign in a SHAP interval tells us if and when a feature has negative or positive impact on classifications. A global analysis of SHAP intervals can also be used to *rank features* according to their general importance.

To determine the SHAP intervals it is necessary to find minimal upper-bounds and maximal lower-bounds for the SHAP score in the uncertainty region. Formulated in terms of thresholds, these problems turn out to be in the class NP[3] for a wide class of classifiers. Furthermore, we establish that this problem becomes NP-complete even for simple models such as *decision trees* (and other classifiers that share some properties with them). Notice that computing SHAP for decision trees can be done in polynomial time under the product and uniform distributions. Actually, this result can be obtained for larger classes of classifiers that include decision trees [3, 27].

We also propose and study three other problems related to the behavior of the SHAP score in the uncertainty region, and obtain the same complexity theoretical results as for the problem of computing the maximum and minimum SHAP score. These problems are: (1) deciding whether there are two different distributions in the uncertainty region such that the SHAP score is positive in one of them and negative in the other one (and therefore there is no certainty on whether the feature contributes positively or negatively to the prediction), (2) deciding if there is some distribution such that the SHAP score is 0 (i.e., if the feature can be considered *irrelevant* in some sense), and (3) deciding if for every distribution in the uncertainty region it holds that a feature x is better ranked than a feature y (i.e., if x *dominates* y).

We remark that the upper bound of NP for all these problems is not evident since they all involve reasoning around polynomial expressions, and in principle we may not have polynomial bounds for the size of the witnesses. Moreover, as we will see further on, the SHAP polynomial cannot even be computed explicitly for most models.

To conclude, we carry out an experimentation to compute these SHAP intervals over a real dataset in order to observe what additional information is provided by the use of the SHAP intervals. We find out that, under the presence of uncertainty, most of the rankings are *sensitive* to the choice of the distribution over the uncertainty region: the ranking may vary depending on the chosen distribution, even when taking into account only the top 3 ranked features. We also study how this sensitivity decreases as the precision of the distribution estimation increases.

Related work. Close to our work, but aiming towards a different direction, we find the problem of *distributional shifts* [8, 18, 15], which in ML occur when the distribution of the training data differs from the data the classifier encounters in a particular application. This discrepancy poses significant challenges, as can lead to decreased performance and unexpected behavior of models.

Also related is the problem of score *robustness* [1, 13]: one can analyze how scores change under small perturbations of an input. In our case, we study uncertainty at the level of the underlying probability distributions.

The work [25] also tries to address uncertainty in the importance of features for local explanations, but does so from a Bayesian perspective: they use a novel sampling procedure to estimate credible intervals around the mean of the feature importance, and derive closed-form expressions for the number of perturbations required to reach explanations within the desired levels of confidence.

Finding optimal intervals under uncertainty as done here, is reminiscent of finding tight ranges for aggregate queries from uncertain databases which are repaired due to violations of integrity constraints [2].

Finally, other lines of work such as [17] aim to understand the uncertainty that arises from approximation errors when computing the Shapley values via a sampling procedure over the feature space. In such contexts, the distribution is usually assumed as given, and therefore these works focus on formalizing a scenario that differs from ours. Moreover, all our analyses and algorithms are based on optimal confidence intervals that arise from exact computation of the Shapley values.

Our contributions. In this work we make the following contributions:

1. We propose a new approach to understand the SHAP score by interpreting it as a polynomial evaluated over an uncertainty region of probability distributions.
2. We analyze at which points of the uncertainty region the maximum and minimum values for SHAP are attained.
3. We establish NP-completeness of deciding if the score of a feature can be larger than a given threshold; we also show NP-completeness for some related problems.

[3] See [12] for a standard introduction into the complexity classes considered in this paper.

4. We provide experimental results showing how SHAP scores can vary over the uncertainty region, and how considering uncertainty makes it possible to define more nuanced rankings of feature importance.

Organization. This paper is structured as follows: In Section 2 we introduce notation and recall basic definitions. In Section 3, we formalize our problems and obtain the first results. Section 4 presents our main complexity results. In Section 5, we describe our experiments and show their outcome. Finally, in Section 6, we make some final remarks and point to open problems. Proofs for all our results can be found in [9].

2 Preliminaries

Let X be a finite set of *features*. An *entity* e over X is a mapping $e : X \to \{0, 1\}$. We denote by $\mathtt{ent}(X)$ the set of all entities over X. Given a subset of features $S \subseteq X$ and an entity e over X, we define the set of entities *consistent with* e on S as:

$$\mathtt{cw}(e, S) := \{e' \in \mathtt{ent}(X) : e'(x) = e(x) \text{ for all } x \in S\}.$$

As already discussed in Section 1, we shall consider *product distributions* as our basic probability distributions over the entity population $\mathtt{ent}(X)$. A product distribution $\mathbb{P}$ over $\mathtt{ent}(X)$ is parameterized by values $(p_x)_{x\in X}$. For every $e \in \mathtt{ent}(X)$ we have:

$$\mathbb{P}(e) = \prod_{x\in X:e(x)=1} p_x \prod_{x\in X:e(x)=0} (1 - p_x).$$

That is, each feature value $e(x)$ is chosen independently with a probability according to p_x ($e(x) = 1$ with probability p_x).

A *(binary) classifier or model* M over X is a mapping $M : \mathtt{ent}(X) \to \{0, 1\}$. We say that M *accepts* e if $M(e) = 1$, otherwise M *rejects* the entity. Let M be a binary classifier and e an entity, both over X. We define the function $\phi_{M,e} : 2^X \to [0, 1]$ as:

$$\phi_{M,e}(S) := \mathbb{E}[M \mid \mathtt{cw}(e, S)].$$

In other words, $\phi_{M,e}(S)$ is the expected value of M conditioned to the event $\mathtt{cw}(e, S)$. More explicitly:

$$\phi_{M,e}(S) = \sum_{e'\in\mathtt{cw}(e,S)} \mathbb{P}(e' \mid \mathtt{cw}(e, S))M(e').$$

A direct calculation shows that the conditional probability $\mathbb{P}(e' \mid \mathtt{cw}(e, S))$ can be written as:

$$\mathbb{P}(e' \mid \mathtt{cw}(e, S)) = \prod_{x\in X\setminus S:e'(x)=1} p_x \prod_{x\in X\setminus S:e'(x)=0} (1 - p_x).$$

The function $\phi_{M,e}$ can be used as the wealth function in the general formula of the Shapley value [23, 21] to obtain the SHAP score of the feature values in e.

Definition 1 (SHAP score)**.** *Given a classifier M over a set of features X, an entity e over X, and a feature $x \in X$, the* SHAP score *of feature x with respect to M and e is*

$$Shap(M, e, x) := \sum_{S\subseteq X\setminus x} c_{|S|} \left(\phi_{M,e}(S \cup \{x\}) - \phi_{M,e}(S)\right),$$

where $c_i := \frac{i!(|X|-i-1)!}{|X|!}$.

Intuitively, the SHAP score intends to measure how the inclusion of x affects the conditional expectation of the prediction. In order to do this, it considers every possible subset $S \subseteq X \setminus \{x\}$ of the features and compares the expectation for the set S against $S \cup \{x\}$. A score close to 1 implies that x heavily leans the classifier M towards acceptance, while a score close to -1 indicates that it leans the prediction towards rejection (note that SHAP always takes values in $[-1, 1]$).

Example 2. *Consider again the model M from Table 1. It can be shown that if $p_x = p_y = \frac{1}{2}$ and $p_z = \frac{3}{4}$ then feature z has SHAP score 0.25 while x and y have score 0.1875. Meanwhile, if $p_z = \frac{1}{4}$ then $Shap(M, e, z) \sim 0.08$ and $Shap(M, e, x) = Shap(M, e, y) \sim 0.15$.*

In practical applications, the exact distribution of entities is generally unknown and subjectively assumed or estimated from data. The previous example shows that the choice of the underlying distribution can have severe effects when establishing the importance of features in the classifications. To overcome these problems, in the next section we formalize the notion of distributional uncertainty and present our framework for reasoning about SHAP scores in that setting.

3 SHAP under Distributional Uncertainty

The general idea of our framework is as follows: we explicitly consider a set that contains the potential distributions. This provides us with what we call the *uncertainty region*. This allows us to reinterpret the SHAP score as a *function* from the uncertainty region to $\mathbb{R}$. We can then analyze the behavior of this function in order to gain concrete insights about the importance of a feature.

Recall that a product distribution is determined by its parameters $(p_x)_{x\in X}$. For convenience, we always assume an implicit ordering on the features. Hence, we can identify our space of probability distributions over $\mathtt{ent}(X)$ with the set $[0, 1]^{|X|}$. In order to define uncertainty regions, it is natural then to consider *hyperrectangles* $\mathcal{I} \subseteq [0, 1]^{|X|}$, i.e., subsets of the form $\mathcal{I} = \times_{x\in X}[a_x, b_x]$. Intuitively, these regions correspond to independently choosing a confidence interval $[a_x, b_x]$ for the unknown probability p_x, for each feature $x \in X$.

Example 3. *Within the setting from Example 1, consider the following uncertainty regions defined by hyperrectangles $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively:*

$$\mathcal{I}_1 := [\frac{1}{3}, \frac{1}{2}] \times [1, 1] \times [\frac{1}{3}, \frac{2}{3}]$$

$$\mathcal{I}_2 := [\frac{1}{2}, \frac{1}{2}] \times [\frac{1}{2}, 1] \times [0, \frac{1}{2}]$$

Notice that the SHAP polynomial in region 1 attains a maximum at $p_x = \frac{1}{3}$, $p_y = 1$, and $p_z = \frac{2}{3}$, where the maximum score is $\frac{8}{27}$. The minimum value, corresponding to the score $\frac{5}{36}$, is attained at $p_x = \frac{1}{2}$, $p_y = 1$, $p_z = \frac{1}{3}$.

For the second region, the maximum is attained at $p_x = \frac{1}{2}$, $p_y = 1$, and $p_z = \frac{1}{2}$, and the maximum value is $\frac{5}{24}$. Similarly, the minimum score 0 is attained whenever $p_z = 0$.

The SHAP score now becomes a function taking probability distributions and returning real values. This is formalized below.

Definition 2 (SHAP polynomial)**.** *Given a classifier M over a set of features X, an entity e over X, and a feature $x \in X$, the* SHAP polynomial *$Shap_{M,e,x}$ is the function from $[0, 1]^X$ to $\mathbb{R}$ mapping each $(p_x)_{x\in X}$ to the SHAP score $Shap(M, e, x)$ using $(p_x)_{x\in X}$ as the underlying product distribution.*

As the name suggests, the SHAP polynomial $Shap_{M,e,x}$ is actually a multivariate polynomial on the variables $(p_x)_{x\in X}$. Moreover, it is a *multilinear* polynomial: it is linear on each of its variables separately (equivalently, no variable occurs at a power of 2 or higher).

Proposition 1. *Given a classifier M over a set of features X, an entity e over X, and a feature $x \in X$, the* SHAP polynomial $Shap_{M,e,x}$ *is a multilinear polynomial on variables $(p_x)_{x\in X}$.*

Proof. Note that $\phi_{M,e}(S)$ is a multilinear polynomial for any subset of features $S \subseteq X$. Since $Shap_{M,e,x}$ is a weighted sum of expressions of the form $\phi_{M,e}(S)$ the result follows by observing that multilinear polynomials are closed with respect to sum and product by constants. □

We can then reason about the importance of features via the analysis of SHAP polynomials. Here we concentrate on the fundamental problems of finding maxima and minima of these polynomials over the uncertainty region. This allows us to determine the *SHAP interval* of a feature, i.e., the set of possible SHAP scores of the feature over the uncertainty region. SHAP intervals may provide useful insights. For instance, obtaining smaller SHAP intervals for a feature suggests its SHAP score is more robust against uncertainty. On the other hand, they can be used to assess the relative importance of features (see Section 5 for more details).

We aim to characterize the complexity of these problems, thus we formulate them as decision problems[4]:

> PROBLEM: REGION-MAX-SHAP
> INPUT: A classifier M, an entity e, a feature x, a hyperrectangle $\mathcal{I}$ and a rational number q.
> OUTPUT: Is there a point $\mathbf{p} \in \mathcal{I}$ such that $Shap_{M,e,x}(\mathbf{p}) \geq q$?.

The problem REGION-MIN-SHAP is defined analogously by requiring $Shap_{M,e,x}(\mathbf{p}) \leq q$. We also consider some related problems:

- REGION-AMBIGUITY: given a classifier M, an entity e, a feature x, and a hyperrectangle $\mathcal{I}$, check whether there are two points $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{I}$ such that $Shap_{M,e,x}(\mathbf{p}_1) > 0$ and $Shap_{M,e,x}(\mathbf{p}_2) < 0$. This problem can be understood as a simpler test for robustness (in comparison to actually computing the SHAP intervals).
- REGION-IRRELEVANCY: given a classifier M, an entity e, a feature x, and a hyperrectangle $\mathcal{I}$, check whether there is a point $\mathbf{p} \in \mathcal{I}$ such that $Shap_{M,e,x}(\mathbf{p}) = 0$. This is the natural adaptation of checking *irrelevancy* of a feature (score equal to 0) to the uncertainty setting.
- FEATURE-DOMINANCE: given a classifier M, an entity e, features x and y, and a hyperrectangle $\mathcal{I}$, check whether x *dominates* y, that is, for all points $\mathbf{p} \in \mathcal{I}$, we have $Shap_{M,e,x}(\mathbf{p}) \geq Shap_{M,e,y}(\mathbf{p})$. The notion of dominance provides a safe way to compare features under uncertainty.

4 Complexity Results

We now present our main technical contributions, namely, we pinpoint the complexity of the problems presented in the previous section.

[4] The encoding of M depends on the class of classifiers considered, while $\mathcal{I}$ is given by listing the rationals a_i, b_i $(1 \leq i \leq n)$.

4.1 Preliminaries on multilinear polynomials

A *vertex* of a hyperrectangle $\mathcal{I} = \times_{i=1}^{n}[a_i, b_i]$ is a point $\mathbf{p} \in \mathcal{I}$ such that $\mathbf{p}_i \in \{a_i, b_i\}$ for each $1 \leq i \leq n$. The following is a well-known fact about multilinear polynomials (see e.g., [14]). For completeness we provide a simple self-contained proof in Appendix A.1 in [9].

Proposition 2. *Let f be a multilinear polynomial over n variables. Let $\mathcal{I} \subseteq [0,1]^n$ be a hyperrectangle. Then the maximum and minimum of f restricted to $\mathcal{I}$ is attained in the vertices of $\mathcal{I}$.*

Proposition 2 yields two algorithmic consequences. On the one hand, it induces an algorithm to find the maximum of f over $\mathcal{I}$ in time $2^n poly(|f|)$: simply evaluate the polynomial on all the vertices, and keep the maximum[5]. On the other hand, it shows that this problem is certifiable: to decide whether f can reach a value as big as q, we just need to guess the corresponding vertex and evaluate it. We show that, within the usual complexity theoretical assumptions, there is no polynomial algorithm to find this maximum (see Appendix A.1 in [9]):

Theorem 3. *Given a multilinear polynomial f, a hyperrectangle $\mathcal{I} = \times_{i=1}^{n}[a_i, b_i]$, and a rational q, the problem of deciding whether there is an $\boldsymbol{x} \in \mathcal{I}$ such that $f(\boldsymbol{x}) \geq q$ is* NP*-complete.*

4.2 Complexity of REGION-MAX-SHAP

It is well-known that computing SHAP scores is hard for general classifiers. For instance, the problem is already #P-hard when considering *Boolean circuits* [3] or *logistic regression models* [27]. For *linear perceptrons*, model counting is intractable [4] and by the results in [3], it follows that SHAP computation for perceptrons is also intractable. On the other hand, computing maxima of SHAP polynomials for a certain class of classifiers is as hard as computing SHAP scores for that class of classifiers: if we consider the hyperrectangle consisting of a single point $\mathcal{I} = \times_{i=1}^{n}[p_i, p_i]$, the maximum of the SHAP polynomial coincides with the SHAP score for the product distribution $(p_i)_{1\leq i\leq n}$. Therefore, we focus on family of classifiers where the SHAP score can be computed in polynomial time. A prominent example is the class of *decomposable and deterministic Boolean circuits*, whose tractable SHAP score computation has been shown recently [3]. This class contains as a special case the well-known class of *decision trees*. For a formal definition see [10, 11]. As a consequence of Proposition 2, we obtain the following complexity upper bound:

Corollary 4. *Let $\mathcal{F}$ be a class of classifiers for which computing the SHAP score for given product distributions can be done in polynomial time. Then REGION-MAX-SHAP is in* NP *for the class $\mathcal{F}$. In particular, REGION-MAX-SHAP is in* NP *for decomposable and deterministic Boolean circuits.*

Next, we show a matching lower bound for REGION-MAX-SHAP. Interestingly, this holds even for decision trees. We stress that the NP-hardness of REGION-MAX-SHAP does not follow directly from Theorem 3: in REGION-MAX-SHAP the multilinear polynomial is given implicitly, and it is by no means obvious how to encode the multilinear polynomials used in the hardness argument of Theorem 3. Instead, we follow a different direction and encode directly the classical problem of VERTEX-COVER.

[5] We assume f is given by listing its non-zero coefficients and their corresponding monomials.

Theorem 5. *The problem REGION-MAX-SHAP is* NP*-hard for decomposable and deterministic Boolean circuits. The result holds even when restricted to decision trees.*

Sketch of the proof. We reduce from the well-known NP-complete problem VERTEX-COVER: given a graph $G = (V, E)$ and $k \geq 1$, decide whether there is a vertex cover[6] of size at most k.

The hardness proof relies on two observations. Firstly, by using $|V|$ features and choosing the hyperrectangle $\mathcal{I} = [0, 1]^{|V|}$ as the uncertainty region, there is a natural bijection $\mathbf{p}(C) = \mathbf{p}^C \in \mathcal{I}$ between the subsets $C \subseteq V$ and the vertices of $\mathcal{I}$ ($v \in C$ iff $p_v = 0$). Secondly, by properly picking the entities accepted by model M, the SHAP polynomial will be

$$Shap_{M,e,x}(\mathbf{p}^C) = -\sum_{\{u,v\}\in E} p_u p_v I_{u,v} - T_{n,\ell}$$

where ℓ is the size of the subset C and the term $T_{n,\ell}$ grows as ℓ grows. The term $I_{u,v}$ is positive and works as a *penalization factor* which is "activated" if $p_u = p_v = 1$, i.e., when the edge uv is uncovered. Furthermore, by adding an extra feature w to the model (and consequently, another probability p_w) we can make this penalization factor arbitrarily big in relation to the *size factor* $T_{n,\ell}$.

We choose the bound q to be $-T_{n,k}$. If C is a vertex cover of size ℓ, then each term $p_u p_v I_{u,v}$ equals 0 and hence $Shap_{M,e,x}(\mathbf{p}^C) = -T_{n,\ell}$. On the other hand, if C is not a vertex cover, then some $p_u p_v I_{u,v}$ is non-zero, and by defining an adequate interval for p_w we can ensure that $-p_v p_v I_{u,v} < q$. Hence, the only way to obtain $Shap_{M,e_0,x_0}(\mathbf{p}^C) \geq q$ is to pick C to be, in the first place, a vertex cover, and secondly, one of size $\ell \leq k$. □

Both Corollary 4 and Theorem 5 also apply to REGION-MIN-SHAP (see Appendix A.3 in [9] for details).

4.3 Related problems

In this section we show some results related to the problems proposed in Section 3. As in the case of REGION-MAX-SHAP they turn out to be NP-complete, even when restricting the input classifiers to be decision trees.

Again, as a consequence of Proposition 2 we obtain the NP membership for REGION-AMBIGUITY and REGION-IRRELEVANCY, and the coNP membership for FEATURE-DOMINANCE.

Corollary 6. *Let $\mathcal{F}$ be a class of classifiers for which computing the SHAP score for given product distributions can be done in polynomial time. Then the problems REGION-AMBIGUITY and REGION-IRRELEVANCY are in* NP *for the class $\mathcal{F}$, while the FEATURE-DOMINANCE is in* coNP[7].

The hardness for these problems follows under the same conditions as Theorem 5.

Theorem 7. *The problems REGION-AMBIGUITY and REGION-IRRELEVANCY are* NP*-hard for decision trees, while FEATURE-DOMINANCE is* coNP*-hard.*

Sketch of the proof. The proof follows the same techniques as the proof for Theorem 5, through a reduction from VERTEX-COVER. For the case of REGION-IRRELEVANCY we devise a model M such that

$$Shap_{M,e,x}(\mathbf{p}^C) = T_{n,k} - \sum_{\{u,v\}\in E} p_u p_v I_{u,v} - T_{n,\ell} \qquad (2)$$

where ℓ is the size of C, $T_{n,\ell}$ corresponds to the *size factor*, and $I_{u,v}$ is the *penalization factor* for uncovered edges. Observe that the first term $T_{n,k}$ does not depend on the set C, and consequently $Shap_{M,e,x}(\mathbf{p}^C) = 0$ if C is a vertex cover of size k. The construction is a bit more complex, and we have to add $2(n-k)$ extra features to those considered in the construction of Theorem 5.

The proof for REGION-AMBIGUITY is obtained by a slight modification of Equation 2 in order to make the SHAP score positive (instead of 0) if C is a vertex cover of size k.

Finally, for the hardness of FEATURE-DOMINANCE we prove that REGION-AMBIGUITY$\leq_p$ $\overline{\text{FEATURE-DOMINANCE}}$[8]. This reduction is achieved by adding a "dummy feature" w that does not affect the prediction of the model. We prove that its SHAP polynomial is constant and equal to 0, and consequently deciding the ambiguity for a feature x is equivalent to deciding the dominance relation between x and w. □

5 Experimentation

As a case study, we are going to use the California Housing dataset [19], a comprehensive collection of housing data within the state of California containing 20,640 samples and eight features[9]. Our choice of dataset relies mainly on the fact that this dataset has already been considered in the context of SHAP score computation [6] and its size and number of features allow us to compute most of the proposed parameters (as the SHAP polynomial itself, for each feature) in a reasonable time. Nonetheless, we recall that the proposed framework can be applied to any dataset, as long as there is some uncertainty on the real distribution of feature values and uncertainty regions can be estimated for it (e.g., via sampling the data as we do here).

Note that, while there is no reason to expect that the probabilities of each feature are independent of each other in the California Housing dataset, we make this assumption in our framework (which is a common one in the literature [26, 24]).

5.1 Objectives

The purpose of this experiment is to use a real dataset to simulate a situation where we have uncertainty over the proportions of each feature in the dataset. We want to derive suitable hyperrectangles representing the distribution uncertainty where our extended concepts of SHAP scores apply, and compare these scores against the usual, point-like SHAP score, in order to reveal cases where these new hypervolume scores are more informative than the traditional scheme. We expect our proposal to be able to detect features whose ranking is vulnerable to small distribution shifts, and aim to study how such sensitivity starts to vanish as we reduce the uncertainty over the distribution (i.e., as the hyperrectangles get smaller).

[6] Recall a vertex cover of $G = (V, E)$ is a subset of the nodes $C \subseteq V$ such that for each edge $\{u, v\} \in E$, either $u \in C$ or $v \in C$.

[7] The proof for FEATURE-DOMINANCE follows by observing that $diff(\mathbf{p}) = Shap_{M,e,x}(\mathbf{p}) - Shap_{M,e,y}(\mathbf{p})$ is again a multilinear polynomial.

[8] If Π is a decision problem, we denote its complement as $\overline{\Pi}$.

[9] The code developed for the experimentation can be found at https://git.exactas.uba.ar/scifuentes/fuzzyshapley.

5.2 Methodology

Preparing dataset Our framework is defined for binary features, and therefore we need to binarize the features from the California Housing dataset. Of the eight features, seven are numerical and one is categorical. To binarize each of the numerical ones, we take the average value across all entities and use it as a threshold. We also binarize the target `median_house_value` using the same strategy. The categorical one is `ocean_proximity`, which describes how close each entity is to the ocean, with the categories being {`INLAND`, `<1H OCEAN`, `NEAR OCEAN`, `NEAR BAY`, `ISLAND`}. The `INLAND` one represents the farthest distance away from the ocean, and we binarize that category to 0, while the other ones are mapped to 1. Finally, we remove the rows with `NaN` values.

The model M We take 70% of the data for training, and keep 30% for testing. As a model we employ the *sklearn* `DecisionTreeClassifier`. For regularization we considered both bounding the depth of the tree by 5 and bounding the minimum samples to split a node by 100. The obtained results were similar for both mechanisms, and therefore we only show here the results for the model regularized by restricting the number of samples required to split a node. The trained model has 80.0% accuracy and 76.5% precision.

Creating the hyperrectangle We assume independence between the different distributions of the features, and ignorance of their true probabilities. Therefore, to obtain an estimation of the probability p_x that a feature x has value 1 for a random entity, we will sample a subset of the available data and compute an empirical average. We vary the number of samples taken in order to simulate situations with different uncertainty.

Let N be the number of samples. Given $0 < p < 1$, we sample $\lceil pN \rceil$ entities 5 times, and for each of these times we compute an empirical average p_x^j for each feature x. We then define the estimation for p_x as the median of $\{p_x^j\}_{1 \leq j \leq 5}$ and compute a deviation σ_x by taking the deviation over the set $\{p_x^j\}_{1 \leq j \leq 5}$. Finally, the hyperrectangle is defined as $\times_x [p_x - \sigma_x, p_x + \sigma_x]$. We experiment with different values for p from 10^{-3} to $\frac{1}{2}$.

Comparing SHAP scores We pick 20 different entities uniformly at random and compute the SHAP score for each of these entities and for each feature at all the vertices of the different hyperrectangles. Instead of using the polynomial-time algorithm for the SHAP value computation over decision trees, we actually compute an algebraic expression for the SHAP polynomial for each pair of entity and feature, simplify it, and then use it to compute any desired SHAP score given any product distribution.

5.3 Results

We recall that the *SHAP interval* of feature x for entity e over the hyperrectangle $\mathcal{I}$ is $[\min_{\mathbf{p} \in \mathcal{I}} Shap_{M,e,x}(\mathbf{p}), \max_{\mathbf{p} \in \mathcal{I}} Shap_{M,e,x}(\mathbf{p})]$.

In Figure 1 we plot the SHAP intervals of all features for one of the entities we used and two sampling percentages. By observing the intervals it is clear that, even in the presence of the uncertainty of the real distribution, as long as it belongs to the hyperrectangle $\mathcal{I}$ the features `median_income` and `ocean_proximity` will be the top 2 in the ranking defined by the SHAP score, for both sampling percentages considered. However, when the sampling percentage is small ($p = 0.4\%$) the SHAP intervals of both features intersect, and therefore it could be the case that their relative ranking actually depends on the chosen distribution inside $\mathcal{I}$.

To decide whether this happens, one should find two points $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{I}$ such that $Shap_{M,e,\texttt{med\_inc}}(\mathbf{p}_1) < Shap_{M,e,\texttt{ocean\_prox}}(\mathbf{p}_1)$ and $Shap_{M,e,\texttt{med\_inc}}(\mathbf{p}_2) > Shap_{M,e,\texttt{ocean\_prox}}(\mathbf{p}_2)$ (i.e., solve the FEATURE-DOMINANCE problem). This can be done by observing that $\mathit{diff}(\mathbf{p}) = Shap_{M,e,\texttt{med\_inc}}(\mathbf{p}) - Shap_{M,e,\texttt{ocean\_prox}}(\mathbf{p})$ is a multilinear polynomial, and therefore its maximum and minimum are attained at the vertices of $\mathcal{I}$. By computing $\mathit{diff}(\mathbf{p})$ on all these points we observed that in 4 of the 256 vertices `median_income` has a bigger SHAP score than `ocean_proximity`, and consequently their relative ranking depends on the chosen underlying distribution.

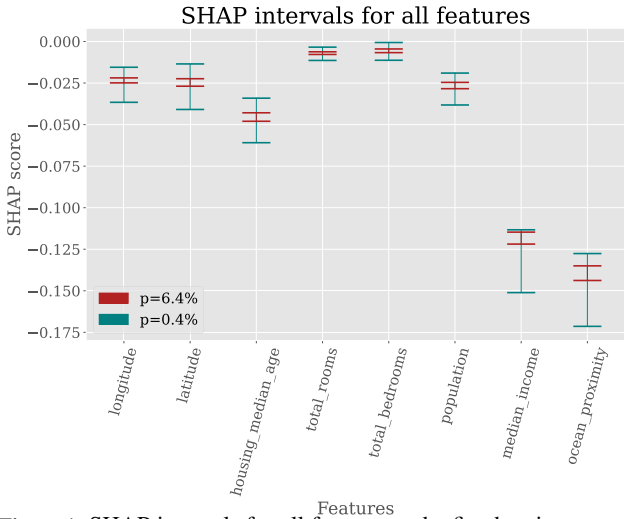

Figure 1: SHAP intervals for all features and a fixed entity, over two different sampling percentages. When $p = 0.4\%$ it is clear that, according to the SHAP score, the features `median_income` and `ocean_proximity` are the two most relevant, but there is an uncertainty on which one of the two is the most important. When the sampling percentage increases to $p = 6.4\%$ the SHAP intervals become disjoint, and we can be certain that `ocean_proximity` is the most relevant feature. Observe that the same kind of uncertainty exists regarding the third ranked feature.

We can also observe in Figure 1 that something similar happens as well for feature `housing_median_age`. When $p = 0.4\%$ its SHAP interval intersects with the intervals of the three features `longitude`, `latitude` and `population`. Nonetheless, by inspecting the difference of the scores at the vertices of $\mathcal{I}$ it can be seen that there is no point in which `housing_median_age` is ranked below the other features (i.e., `housing_median_age` dominates all three features). Meanwhile, if we compare two of the other features such as `longitude` and `latitude`, we observe that they have a relevant intersection even when $p = 6.4\%$, and in 179 out of the 256 vertices `latitude` is ranked below `longitude`.

In Figure 2 we can see how the SHAP intervals shrink as the sampling percentage increases. For all sampling percentages above 6.4% we can say with certainty that `ocean_proximity` is ranked above `median_income`. This behavior is natural since the size of the hyperrectangles is getting smaller, because the deviation of the empirical samples p_x^i tends to get smaller as the sampling percentage increases.

Finally, in Figure 3 we plot the number of entities whose ranking depends on the chosen distribution, for each sampling percentage, and restricting ourselves to some subset of the ranking. We found out that for 10 of the 20 entities the complete ranking defined by the

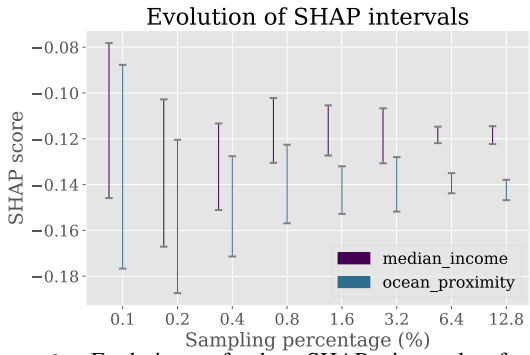

Figure 2: Evolution of the SHAP intervals for features `median_income` and `ocean_proximity` considering the same entity as in Figure 1 and the different sampling percentages.

SHAP score depends on the chosen distribution even when the sample percentage is as big as $p = 12.8\%$ (recall that due to the way we built our hyperrectangles, we are sampling $5p$ times the dataset, and therefore if $p > 10\%$ we might be inspecting more than half of the data points). If we are only concerned with the top three ranked features, we can observe that even when $p = 6.4\%$ there are still 4 entities for which the top 3 ranking is sensitive to the selected distribution.

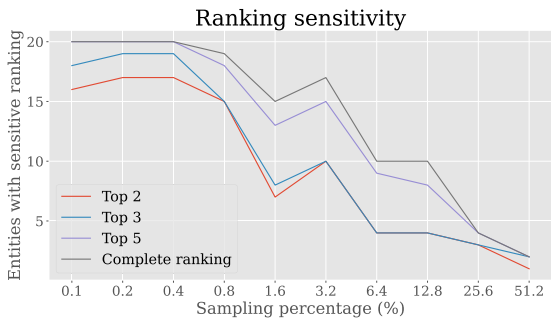

Figure 3: Number of entities whose feature ranking may vary depending on the chosen distribution inside the uncertainty hyperrectangle, for each sampling percentage. The `Top` k line indicates whether there was a change in the ranking of the top k features, ignoring changes in the rest of the ranking. As expected, sensitivity of the ranking is more common when the sampling percentage is smaller.

6 Conclusions

We have analyzed how SHAP scores for classification problems depend on the underlying distribution over the entity population when the distribution varies over a given uncertainty region. As a first and important step, we focused on product distributions and hyperrectangles as uncertainty regions, and obtained algorithmic and complexity results for fundamental problems related to how SHAP scores vary under these conditions. As a proof of concept, we showed through experimentation that considering uncertainty regions has an impact on feature rankings for a non-negligible percentage of the entities, and that the solutions to our proposed problems can provide insight on the relative rankings even in the presence of uncertainty.

Feature Independence. We stress that from a complexity point of view, it is natural to start with product distributions. As shown in [27], the computation of SHAP scores becomes intractable for trivial classifiers as soon as we consider simple non-product distributions such as naive Bayes distributions. As a consequence, computing maxima of SHAP polynomials is trivially intractable in this setting. As we discussed at the beginning of Section 4.2, we focused on the cases for which computing SHAP scores is tractable, since this gives us the possibility to also obtain tractability for computing maxima of SHAP polynomials. We considered the prominent case of decomposable and deterministic Boolean circuits under product distributions. This case has been shown to be tractable in [27] and [3] (in [27] product distributions are referred to as fully-factorized distributions).

For our distributional shift analysis, considering more general distributions on the entity population would still require specifying a class of them and their regions of variation. A natural next step in this direction, that would build on our work, consists in imposing additional *domain knowledge* on the product distribution, leading, in particular, to certain dependencies among features. For example, a constraint specifying that "house lots located at the seaside have a price about $2M". More generally, a conjunction φ of such constraints is associated to an event $E^{\varphi} \subseteq \mathtt{ent}(X)$ containing all entities that satisfy it. Conditioning on this event leads to a new distribution $\mathbb{P}'$ defined by $\mathbb{P}'(A) := \mathbb{P}(A|E^{\varphi})$, for $A \subseteq \mathtt{ent}(X)$. The shift analysis would be done on the new entity space $\langle \mathtt{ent}(X), \mathbb{P}' \rangle$, where (in general) features will not be fully independent anymore. This would be done, without having to start from scratch with $\mathbb{P}'$, by taking advantage of our previous analysis of the original product distribution $\mathbb{P}$.

It is worth mentioning that imposing and exploiting *domain semantics* when defining and computing explanation scores is interesting in its own right (see [5] in relation to the RESP score). The problem of using domain knowledge in Explainable AI has been scarcely investigated in the literature.

Explanatory Robustness. It would be interesting to know how a local perturbation of a given distribution in the uncertainty region affects the SHAP scores, or their rankings. This would be a proper *robustness analysis* with respect to the distribution (as opposed to how SHAP scores vary in a region). Of course, this is a different problem from the most common one of robustness with respect to the perturbation of an input entity (see e.g., [1, 13]).

There are several other problems left open by our work. The inclusion of non-binary features and labels is a natural next step. It would also be interesting to analyze others proposed scores (e.g., LIME [20], RESP [7], Kernel-SHAP [16]) in the setting of distributional uncertainty.

Acknowledgments

This research has been supported by an STIC-AmSud program, project AMSUD210006. We appreciate the code made available by Jorge E. León (UAI, Chile) for SHAP computation with the same dataset used in our work. Bertossi has been supported by NSERC-DG 2023-04650, the Millennium Institute for Foundational Research on Data (IMFD, Chile), and CENIA (Basal ANID FB210017, Chile). Romero is funded by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Pardal is funded by DFG VI 1045-1/1. Abriola, Cifuentes and Pardal are funded by FONCyT, ANPCyT and CONICET, in the context of the projects PICT 01481 and PIP 11220200101408CO. Abriola is additionally funded by the project PIBAA 28720210100188CO. Martinez is partially funded under the Spanish project PID2022-139835NB-C21 funded by MCIN/AEI/10.13039/501100011033, PIE 20235AT010 and iTrust (PCI2022-135010-2). Pardal was supported by the DFG grant VI 1045-1/1.

References

[1] D. Alvarez-Melis and T. Jaakkola. On the Robustness of Interpretability Methods. *ArXiv preprint: 1806.08049*, 2018.

[2] M. Arenas, L. Bertossi, J. Chomicki, X. He, V. Raghavan, and J. Spinrad. Scalar Aggregation in Inconsistent Databases. *Theoretical Computer Science*, 296(3):405–434, 2003.

[3] M. Arenas, P. Barcelo, L. Bertossi, and M. Monet. On the Complexity of SHAP-Score- Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results. *J. Machine Learning Research*, 24(63):1–58, 2023.

[4] P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux. Model Interpretability through the Lens of Computational Complexity. *Advances in Neural Information Processing Systems*, 2020.

[5] L. Bertossi. Declarative Approaches to Counterfactual Explanations for Classification. *Theory and Practice of Logic Programming*, 23(3):559–593, 2023.

[6] L. Bertossi and J. E. León. Efficient Computation of Shap Explanation Scores for Neural Network Classifiers via Knowledge Compilation. In *Proc. European Conference on Logics in Artificial Intelligence (JELIA)*, pages 49–64. Springer LNCS 14281, 2023.

[7] L. Bertossi, J. Li, M. Schleich, D. Suciu, and Z. Vagena. Causality-based Explanation of Classification Outcomes. *Proc. 4th SIGMOD Int. WS on Data Management for End-to-End Machine Learning*, pages 70–81, 2020.

[8] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8): 832, 2019.

[9] S. Cifuentes, L. Bertossi, N. Pardal, S. Abriola, M. V. Martinez, and M. Romero. The Distributional Uncertainty of the SHAP score in Explainable Machine Learning. *arXiv preprint 2401.12731*, 2024.

[10] A. 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.

[11] A. Darwiche and P. Marquis. A Knowledge Compilation Map. *J. Artif. Int. Res.*, 17(1):229–264, 2002. ISSN 1076-9757.

[12] M. R. Garey and D. S. Johnson. *Computers and Intractability*, volume 174. Freeman, San Francisco, 1979.

[13] X. Huang and J. Marques-Silva. From Robustness to Explainability and Back Again. *ArXiv preprint: 2306.03048*, 2023.

[14] C. Laneve, T. A. Lascu, and V. Sordoni. The Interval Analysis of Multilinear Expressions. *Electronic Notes in Theoretical Computer Science*, 267(2):43–53, 2010.

[15] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23 (1):18, 2020.

[16] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[17] L. Merrick and A. Taly. The Explanation Game: Explaining Machine Learning Models using Shapley Values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 17–38. Springer, 2020.

[18] C. Molnar. *Interpretable Machine Learning*. 2020.

[19] Nugent, C. California Housing Prices. https://www.kaggle.com/datasets/camnugent/california-housing-prices, 2018.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of any Classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[21] A. E. Roth, editor. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge Univ. Press, 1988.

[22] S. Lundberg et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[23] L. S. Shapley. A Value for n-Person Games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[24] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features through Propagating Activation Differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[25] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju. Reliable Post Hoc Explanations: Modeling Uncertainty in Explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404, 2021.

[26] E. Štrumbelj and I. Kononenko. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*, 41:647–665, 2014.

[27] G. Van Den Broeck, A. Lykov, M. Schleich, and D. Suciu. On the Tractability of SHAP Explanations. *J. Artif. Intell. Res.*, 74:851–886, 2022.