Building a Bridge Between the Shapley Value, Prime Implicants and Counterfactual Explanations: a Theoretical Analysis

Hénoïk Willot[1]

[1]Sorbonne Université, CNRS, LIP6 , Paris, France

Abstract

Since the introduction of SHAP, many methods adapted the idea of using the Shapley value formula as a way to measure the importance of a feature in the decision of a model. However, computing this value theoretically imply to relearn the model with and without the feature for all subsets of features. To avoid this, the methods rely on sampling to compute an expectation to measure a mean impact of a feature. In this paper we investigate theoretically if we could, and under which conditions, build and use a textbook cooperative game with the help of other explanation techniques, Prime Implicants and Counterfactual explanations, for computing Shapley values for interpretability. We propose detailed and iterative analysis of the number of remaining explanations during the enumeration of the explanations, alongside the computation of the Shapley value and its upper and lower bounds.

1 INTRODUCTION

The Shapley value is originally a value function coming from collaborative game theory Shapley and Shubik [1954]. In this setting, a set of agents are collaborating together, forming coalitions, and the purpose of the Shapley value is to redistribute to each individual agent some portion of the total gain, obtained by the coalition of all agents, based on the agents marginal contribution to the formation of the whole coalition. Based on this purpose, the Shapley value has been vastly used as a power index to determine which features were important for the decision made by a model on an observation Rozemberczki et al. [2022].

Most works in the literature based on SHAP Lundberg and Lee [2017] resolves the issue of computing the game value through sampling. It is well known that computing the Shapley value is an hard task Deng and Papadimitriou [1994] and requires structural constraints Bilbao et al. [2000], Conitzer and Sandholm [2004], Ieong and Shoham [2005], Michalak et al. [2013] or approximation Fatima et al. [2007], Touati et al. [2021], Gemp et al. [2024], Kempinski and Kachman [2025] to be computed. The set function of the collaborative game would require, in an explainability setting, to learn the classifier with any subset of features to capture the marginal impact of using the feature in the prediction, which is infeasible. Instead, with the Shapley value formula, an expectation is distributed between features, obtained through sampling w.r.t a law to be defined. The consequence is that a feature could be assigned a negative value, which is impossible in standard cooperative game theory. Also, depending on the expectation formulation Létoffé et al. [2025], the total amount to be distributed can be more or less normalized which can make the interpretation of the score more complex. Many criticism were raised w.r.t different angles, such as the way and limits of sampling Janzing et al. [2020], Slack et al. [2020], the definition of the expectation function Létoffé et al. [2025], the relevance of the axioms in the interpretability context Fryer et al. [2021] or misleading information Huang and Marques-Silva [2024].

We believe that replacing a collaborative game with some expectation, computed through sampling by an ad hoc probability distribution, certainly allows computations, but also introduces issues with the interpretability of the value, its robustness and its faithfulness, which are important features for building trustworthy models. In contrast, we intend to use well defined cooperative games, building on the notions of Prime Implicants Darwiche and Marquis [2021] and Counterfactual explanations Ignatiev et al. [2020]. These explanations are not unique and enumeration algorithms exists de Colnet and Marquis [2022], Marques-Silva et al. [2020]. We then have multiple objectives in this paper:

- defining and analyzing the properties of our games,
- see how we can rely only on the enumeration of explanations and iteratively compute our values,
- being able to quantify how many explanations can still

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:7387–7402.

theoretically be enumerated,

- compute the Shapley value and bounds on its possible variations w.r.t future steps in the enumeration.

The structure is as follows. We introduce our mathematical notations and the expression of the Shapley value in Section 2. We then define our collaborative games in Section 3. In Section 4 we study the compact form of our game through the set of Prime Implicants or Counterfactual explanations and find the maximum theoretical number of remaining explanations. Then in Section 5 we study the computation of the Shapley value, and lower and upper bounds in Section 6. Finally we conclude and criticize our approach in Section 7.

2 MATHEMATICAL NOTATIONS

In this paper, we are interested in studying the output predictions of some model, noted M, that must predict a categorical variable Y taking values in $\mathcal{Y}$, associated to an observation $\mathbf{x}$ defined over n input variables, or features, $\mathrm{X}_1, \ldots, \mathrm{X}_n$, with their set noted $\mathcal{N} = \{1, \ldots, n\}$, and which respectively takes values in discrete domains $\mathcal{X}_i$. Let $M(\mathbf{x}) = \mathbf{y}$ denote the fact that $\mathbf{y}$ is the prediction by model M over the input sample $\mathbf{x}$.

When considering a subset $\mathcal{S} \subseteq \mathcal{N}$ of dimensions, let $\mathcal{X}_\mathcal{S} = \times^{i\in\mathcal{S}}\mathcal{X}_i$ denote the corresponding domain, and $x_\mathcal{S}$ the values of a vector on this sub-domain. By abusing the notation, $x_\mathcal{S}$ also corresponds to the restriction of vector x on $\mathcal{S}$. We will also denote by $-\mathcal{S} := \mathcal{N} \setminus \mathcal{S}$ all dimensions not in $\mathcal{S}$, with $\mathcal{X}_{-\mathcal{S}}, x_{-\mathcal{S}}$ following the same conventions as $\mathcal{X}_\mathcal{S}, x_\mathcal{S}$. Let $(x_\mathcal{S}, x'_{-\mathcal{S}})$ denote the concatenation of two vectors whose values are given for different features.

The Shapley value of feature i, noted ϕ_i, is defined on a game, represented by a set value function $\nu : 2^\mathcal{N} \to \mathbb{R}$, and is computed with the formulae :

$$c(\mathcal{S}) = \frac{|\mathcal{S}|! \cdot (|\mathcal{N}| - |\mathcal{S}| - 1)!}{|\mathcal{N}|!} \quad (1)$$

$$\phi_i(\nu) = \sum_{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}} c(\mathcal{S}) \cdot (\nu(\mathcal{S}\cup\{i\}) - \nu(\mathcal{S})) \quad (2)$$

The wide use of the Shapley value is also based on its axiomatic properties Rapoport [2001]. Indeed the Shapley value is the only function which verifies :

1. Efficiency: $\sum_{i\in\mathcal{N}} \phi_i(\nu) = \nu(\mathcal{N})$;
2. Symmetry: $\forall \mathcal{S} \subseteq \mathcal{N} \setminus \{i, j\},\ \nu(\mathcal{S}\cup\{i\}) = \nu(\mathcal{S}\cup\{j\}) \Rightarrow \phi_i(\nu) = \phi_j(\nu)$;
3. Null player: $\forall \mathcal{S} \subseteq \mathcal{N} \setminus \{i\},\ \nu(\mathcal{S}) = \nu(\mathcal{S}\cup\{i\}) \Rightarrow \phi_i(\nu) = 0$;
4. Linearity: $\phi_i(\nu^1 + \nu^2) = \phi_i(\nu^1) + \phi_i(\nu^2)$.

3 PREDICTION SEEN AS A COLLABORATIVE GAME

We propose two new definitions of collaborative games, which could be named "prediction games":

Definition 1 (Prediction Game)**.**

$$\nu^i(\mathcal{S}) = \begin{cases} 1 & \textit{if } \forall x_{-\mathcal{S}} \in \mathcal{X}_{-\mathcal{S}}\ M((\mathbf{x}_\mathcal{S}, x_{-\mathcal{S}})) = \mathbf{y}, \\ 0 & \textit{otherwise} \end{cases} \quad (3)$$

i.e. the coalition receives 1 if and only if fixing the values of the features in the coalition to their observed values in $\mathbf{x}$ *is sufficient to ensure the prediction* $\mathbf{y}$*, no matter the values* $x_{-\mathcal{S}} \in \mathcal{X}_{-\mathcal{S}}$ *taken by the features outside the coalition.*

$$\nu^c(\mathcal{S}) = \begin{cases} 1 & \textit{if } \exists x_\mathcal{S} \in \mathcal{X}_\mathcal{S}\ M((x_\mathcal{S}, \mathbf{x}_{-\mathcal{S}})) \neq \mathbf{y}, \\ 0 & \textit{otherwise} \end{cases} \quad (4)$$

i.e. the coalition receives 1 if and only if changing the values of the features inside the coalition to values different from the observation $\mathbf{x}$ *can be sufficient to ensure a change from the prediction* $\mathbf{y}$*, while the values of the features outside the coalition are kept fixed to the observation.*

In the following, we will use interchangeably the terms of feature, player and voter, and interchangeably the terms of coalition and subset of features, as they represent the same concepts in the context of this paper.

3.1 LINKS WITH IMPLICANTS AND COUNTERFACTUAL EXPLANATIONS

The definition of our prediction games through Equations (3) and (4) are not chosen randomly. Indeed, their definitions capture the standard definitions of Prime Implicant explanations and Counterfactual explanations.

Definition 2 (Prime Implicant Explanation)**.** *A subset* $\mathcal{S} \subseteq \mathcal{N}$ *of features is a Prime Implicant explanation (PIX) of* $M(\mathbf{x})$ *iif* $\forall x_{-\mathcal{S}} \in \mathcal{X}_{-\mathcal{S}}\ M((\mathbf{x}_\mathcal{S}, x_{-\mathcal{S}})) = \mathbf{y}$ *and which is subset minimal* $\forall i \in \mathcal{S}\ \exists x_{-\mathcal{S}\cup\{i\}} \in \mathcal{X}_{-\mathcal{S}\cup\{i\}}\ M((\mathbf{x}_{\mathcal{S}\setminus\{i\}}, x_{-\mathcal{S}\cup\{i\}})) \neq \mathbf{y}$.

Definition 3 (Counterfactual Explanation)**.** *A subset* $\mathcal{S} \subseteq \mathcal{N}$ *of features is a subset minimal Counterfactual explanation (CfX) of* $M(\mathbf{x})$ *iif* $\exists x_\mathcal{S} \in \mathcal{X}_\mathcal{S}\ M((x_\mathcal{S}, \mathbf{x}_{-\mathcal{S}})) \neq \mathbf{y}$ *and which is subset minimal* $\forall i \in \mathcal{S}\ \forall x_{\mathcal{S}\setminus\{i\}} \in \mathcal{X}_{\mathcal{S}\setminus\{i\}}\ M((x_{\mathcal{S}\setminus\{i\}}, \mathbf{x}_{-\mathcal{S}\cup\{i\}})) = \mathbf{y}$.

It can be noted that PIXs and CfXs are specific types of coalitions with a valuation of 1, as they have the property that all their sub-coalitions have a valuation of 0, but we will come back to this property later.

These definitions are relevant for two main reasons. First, using the Shapley value as an explanation of which feature had an impact on the decision is intrinsically linked to the ability to say "this feature (and its observed value), by its presence or absence in the observation, changes the prediction". In a sense, it is linked to the "frequency" of each feature to be used as a sufficient reason to change or keep the decision. Then, building on the advances in model compilation Shih et al. [2019], Audemard et al. [2020], de Colnet and Marquis [2023] and explanation enumeration Marques-Silva et al. [2020], de Colnet and Marquis [2022], Audemard et al. [2024], some iterative computation of the Shapley value can be imagined. The results in this paper are general and model agnostic, up to the ability to compute the explanations through the underlying generation process.

3.2 LINKS WITH VOTING GAMES

The fact that ν^i and ν^c are functions in $2^{\mathcal{N}} \rightarrow \{0,1\}$ recalls the notion of simple games (or voting games), well studied in social choice Brandt et al. [2016]. The properties induced by the Implicants and Counterfactual explanations turn indeed ν^i and ν^c into simple games, as they are monotone, but without other structural properties in general.

Proposition 1 (Monotonicity). *ν^i and ν^c are monotone, i.e. $\nu(C) \leq \nu(D)\, \forall C, D \subseteq \mathcal{N}, C \subseteq D$.*

Proof. Obvious if $\nu^i(C) = 0$. It follows from the definition of an Implicant that $\nu^i(C) = 1 \Rightarrow \nu^i(D) = 1$ as $\forall x_{-\mathcal{S}} \in \mathcal{X}_{-\mathcal{S}}\ M((\mathbf{x}_{\mathcal{S}}, x_{-\mathcal{S}})) = \mathbf{y} \Rightarrow \forall x_{-\mathcal{S}\setminus\{i\}} \in \mathcal{X}_{-\mathcal{S}\setminus\{i\}}\ M((\mathbf{x}_{\mathcal{S}\cup\{i\}}, x_{-\mathcal{S}\setminus\{i\}})) = \mathbf{y}$. The same property holds for Counterfactual explanations and ν^c. □

Because ν^i and ν^c are monotone and that the classifier is not vacuous ($\exists x \in \mathcal{X}\ M(x) \neq \mathbf{y}$, *i.e.* $\nu^i(\emptyset) = 0$ and $\nu^c(\mathcal{N}) = 1$, and $M(\mathbf{x}) = \mathbf{y}$, *i.e.* $\nu^i(\mathcal{N}) = 1$ and $\nu^c(\emptyset) = 0$), making the games respect unanimity, they are indeed simple games. However, in the general case, there are no other structural properties. For instance, the union and intersection properties of PIXs and CfXs makes the game neither sub or super modular (see Appendix A).

To match the notions of voting games, a PIX or a CfX is a *minimum winning coalition*, and regular Implicants and Counterfactual explanations are *winning coalitions*. The computation of the Shapley value then only depend on those winning coalitions, as they are the only examples in the game where $\nu(C) - \nu(C \setminus \{i\}) = 1$ for at least some i, represented by the notion of pivotal feature.

Definition 4 (Pivotal Feature). *A feature i is pivotal (or swing) for a coalition C when i turns the coalition from a losing to a winning one, i.e., $\nu(C) = 0$ and $\nu(C \cup \{i\}) = 1$.*

There is then an equivalence between the concepts of *irrelevant feature* Audemard et al. [2021], Létoffé et al. [2025] with the *dummy player* of the game, *i.e.* a feature irrelevant in PIXs and CfXs (never with a useful impact on the decision swap) is assigned a Shapley value of 0; and between the concepts of *mandatory feature* Audemard et al. [2021] and *veto player*, *i.e.* a feature present in all PIXs/CfXs is present in all winning coalitions of the game.

It is also finally possible to give an interpretation to the Shapley value in the context of explanability for a feature i: it is the percentage of all feature permutations in which feature i is pivotal (*i.e.* changes the decision) Shapley and Shubik [1954]. It will also follow that the caveats identified for voting games will also apply to the Shapley value for explanations, which will be discussed in Section 7.

4 COMPUTING THE PREDICTION GAME

Now that we studied the possibilities of our prediction games ν^i and ν^c, we will now handle the less easy task of computing them from $M(\mathbf{x})$. This task is hard, as we do not have only to deal with the exponential size of the number of coalitions of $\mathcal{N}$, but also with the need to ensure the robustness of the decision made on the coalition.

However, there is no need to sample all coalitions, only the winning coalitions, e.g. with coalitions bc and bcd in Figure 1d. We will show in the remainder of this article that all our objectives, namely knowing the number of remaining explanations, computing the Shapley value of a prediction game and bounding its value, can be done solely from the minimum winning coalitions, *i.e.* the list of PIXs or CfXs.

Let's now define the compact version of ν^i and ν^c.

Definition 5 (Compact Games). *$\mu \subseteq 2^{\mathcal{N}}$ is the compact representation of the prediction game ν if and only if :*

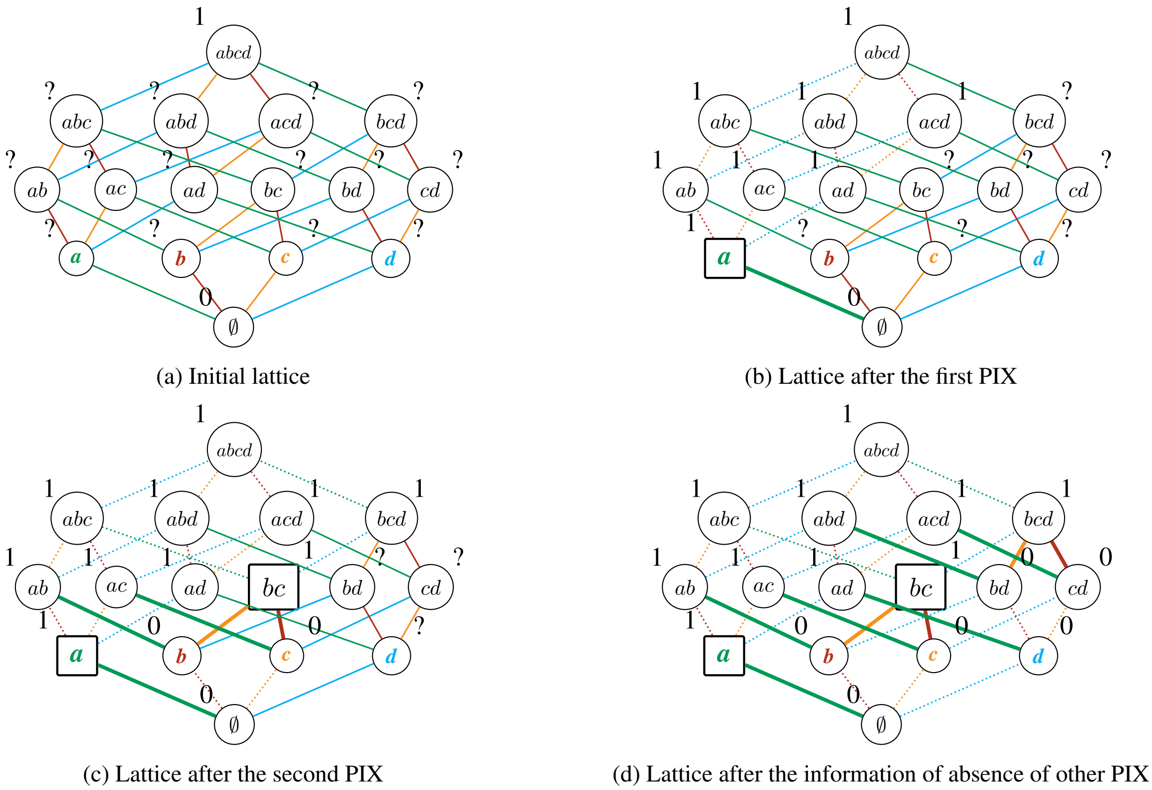
$$\mu = \{\mathcal{S} \in 2^{\mathcal{N}} \mid \nu(\mathcal{S}) = 1, \forall i \in \mathcal{S},\ \nu(\mathcal{S} \setminus \{i\}) = 0\} \quad (5)$$

μ^i represents then the set of PIXs and μ^c the set of CfXs of $M(\mathbf{x})$. These sets are sufficient to compute back the game, as $\nu(C) = 1$ if $\exists \mathcal{S} \in \mu \mid \mathcal{S} \subseteq C$, 0 otherwise. Let us denote by ν^μ the game re-obtained from the compact game μ (hence equal to ν). Building on the literature on model compilation, these sets can be enumerated for a large variety of models.

4.1 COMPACT GAMES STRUCTURE

The compact game (CG), seen as the set of PIXs or CfXs of the decision $M(\mathbf{x})$, has a maximum theoretical size that can be inferred from the structural property of the monotonic set function.

Theorem 1 (Maximum Size of CGs). *The maximum size of a CG $\mu \subseteq \mathcal{N}$ based on the prediction $M(\mathbf{x})$ is $\binom{n}{\lfloor n/2 \rfloor}$.*



(a) Initial lattice

(b) Lattice after the first PIX

(c) Lattice after the second PIX

(d) Lattice after the information of absence of other PIX

Figure 1: Evolution of the game valuation on the Boolean lattice of subsets with the PIX enumeration.
Edges of color green, red, yellow, blue appear in the computation of the Shapley value of resp. a, b, c, d. A bold (resp. dotted) edge is definitively added (resp. invalidated) to the Shapley value. Box-shaped nodes are the enumerated PIX coalitions and the value next to nodes is the game value, with ? if the value is not definitive w.r.t PIXs.

Proof. By definition of a minimum winning coalition $\mathcal{S}$, all the sub-coalitions $\mathcal{S}' \subset \mathcal{S}$ verify $\nu(\mathcal{S}') = 0$, and all the super-coalitions $\mathcal{S}" \supset \mathcal{S}$ verify $\nu(\mathcal{S}") = 1$. Therefore the coalitions in μ form an antichain in the Boolean lattice of subsets and from Sperner's theorem the maximum sized antichain is of length $\binom{n}{\lfloor n/2 \rfloor}$. $\square$

This limit case corresponds to a model M where observing (resp. changing) $\lfloor n/2 \rfloor$ features of $\mathbf{x}$ is sufficient to take (resp. change) the prediction $M(\mathbf{x})$. The explanations correspond then to all the $\lfloor n/2 \rfloor$ sized coalitions[1]. It is also reminescent of the result from Deng and Papadimitriou [1994] as a special case of weighted majority voting. It follows that this *a priori* number of Prime Impliquant or Counterfactual explanations is exponential, bounded by $2^n/\sqrt{n}$, using the bound of the central binomial coefficient.

However, this theoretical *a priori* size and more importantly the trivial identification of which coalition size contains the largest antichain do not hold when the enumeration of the minimum winning coalitions of the game begins. For instance, starting from the initial lattice in Figure 1a, if the PIXs a and b had been enumerated, the maximum length antichain would be built by adding c and d, coalitions of length $1 \neq \lfloor 4/2 \rfloor$, while the only coalition of length 2 available would have been cd. Therefore, keeping track of the maximum remaining number of PIXs/CfXs requires some iterative updates and monitoring, as the number of coalitions of size $\lfloor n/2 \rfloor$ which are not a subset or superset of some enumerated explanation might not give the maximum sized antichain[2].

4.2 PARTIAL COMPACT GAMES

To handle this counting problem during the enumeration, we begin by formalizing the notion of partial compact game.

Definition 6 (Partial Compact Games)**.** *$\mu \subseteq 2^{\mathcal{N}}$ is a partial compact game (PCG) of the CG $\mu' \subseteq 2^{\mathcal{N}}$ if $\mu \subset \mu'$. μ is partial in the sense that not all components of μ' have been observed and many completions are available at that stage. Let $\mathfrak{C}(\mu) \in 2^{2^{\mathcal{N}}}$ define the possible completions of μ:*

$$\mathfrak{C}(\mu) = \{\mu' \subseteq 2^{\mathcal{N}} \mid \mu \subseteq \mu', \forall C, C' \in \mu' \quad C \neq C' \Rightarrow C \nsubseteq C' \text{ and } C' \nsubseteq C\} \tag{6}$$

[1] It also holds for all the $\lceil n/2 \rceil$ sized coalitions when n is odd.

[2] In Figure 1c adding bd and cd still gives the longest antichain.

Finding the maximum number of PIXs or CfXs remaining after the enumeration of μ corresponds to finding the size of $\mu' = \arg\max_{\mu' \in \mathfrak{C}(\mu)} |\mu'|$. As introduced with 4 variables, the layer with the largest number of coalitions might not stay the same during the enumeration. This is due to the fact that only some part of the lattice remains unconnected to the already known minimum winning subsets, as represented in Figure 1 by the question marks in place of the evaluation of the coalitions.

Definition 7 (Unlabeled Set)**.** *Let μ be a PCG. The unlabeled set of μ, noted $\mathfrak{U}(\mu)$ corresponds to the set of the subsets $C \subseteq \mathcal{N}$ for which the value of the game is neither forced to 1 nor to 0, i.e.:*

$$\mathfrak{U}(\mu) = \{C \subseteq \mathcal{N} \mid \forall \mathcal{S} \in \mu, \mathcal{S} \not\subseteq C \text{ and } C \not\subseteq \mathcal{S}\} \qquad (7)$$

It follows from Definition 7 that $\mathfrak{C}(\mu)$ contains all PCG obtained by union of μ with elements of the unlabeled set such that the antichain property is kept, *i.e.* :

$$\begin{aligned}\mathfrak{C}(\mu) = \{&\mu \cup \mu' \mid \mu' \subseteq \mathfrak{U}(\mu),\\ &\forall C, C' \in \mu', C \neq C' \Rightarrow C \not\subseteq C' \text{ and } C' \not\subseteq C\} \qquad (8)\end{aligned}$$

Our notions of PCG and unlabeled set are directly linked to incomplete (or partial) cooperative games from game theory Willson [1993], Masuya and Inuiguchi [2016], Masuya [2021], Cerný and Grabisch [2024], Úradník et al. [2024]. Yet, they differ as our is centered on voting games and minimal winning coalitions, with monotonicity as only structural property (while superadditivity or convexity were studied in these approaches).

Maximum Size of a PCG The maximum number of remaining PIXs/CfXs is then the length of the maximum antichain in the unlabeled set. Saying so is not enough, as its structure might not be regular at all. The evolution of unseen coalitions in all layers of the lattice has to be tracked, the length of the largest one being the maximum number of remaining PIXs/CfXs. Storing all the coalitions in the unlabeled set $\mathfrak{U}(\mu)$, is infeasible, yet, as only the number of such coalitions for each size of subset is required, it is possible to directly count them from the coalitions in μ.

Theorem 2 (Number of Possible Minimum Winning Coalitions of Size k)**.** *Let μ^{k+}, μ^{k-} and μ^{k}, be subsets of PCG μ, forming a partition, such that they contains the coalitions of size respectively strictly greater, strictly smaller and equal to k. The number of possible minimum winning coalitions of size k for PCG μ, noted $\mathfrak{P}(\mu, k)$, is defined by:*

$$\begin{aligned}\mathfrak{P}(\mu, k) = &\binom{n}{k} + \sum_{\substack{\mathbb{S} \subseteq \mu^{k-}\\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}\\ &- |\mu^k| + \sum_{\substack{\mathbb{S} \subseteq \mu^{k+}\\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k} \qquad (9)\end{aligned}$$

Proof. Counting down from the total number of coalitions of size k, minus the number of already enumerated coalitions of size k, the number of coalitions C s.t. $\mathcal{S} \subseteq C$ or $C \subseteq \mathcal{S}$ for some $\mathcal{S} \in \mu$ (*i.e.* C does not belong to $\mathfrak{U}(\mu)$) is to be removed. The first case $\mathcal{S} \subseteq C$ is obtained by projecting cardinally-smaller coalitions to size k, hence choosing $k - |\mathcal{S}|$ features (as the features of $\mathcal{S}$ in C are fixed) from the remaining $n - |\mathcal{S}|$ features. However, some coalitions might be counted several times, as they are superset of more than one minimum winning coalition, hence the inclusion–exclusion principle has to be applied between coalitions of μ^{k-}. The second case comes from the same reasoning by projecting cardinally-bigger coalitions to size k, hence choosing k features from $\mathcal{S}$. Crossing terms between μ^{k+} and μ^{k-} do not have to be considered as $\mathcal{S}^{k-} \subseteq C \subseteq \mathcal{S}^{k+}$ with $\mathcal{S}^{k-} \in \mu^{k-}$ and $\mathcal{S}^{k+} \in \mu^{k+}$ does not exist, as $\mathcal{S}^{k-} \not\subseteq \mathcal{S}^{k+}$ by definition. ☐

Lemma 3 (Maximum Size from a PCG)**.**

$$\max_{\mu' \in \mathfrak{C}(\mu)} |\mu'| = |\mu| + \max_{k \in \{1,\ldots,n-1\}} \mathfrak{P}(\mu, k)$$

Iterative Computation From the expression in Equation (9), an iterative algorithm can be designed by updating the previous value of some size k by the changes caused by the new minimum winning coalition and its intersections/unions with the previous PCG.

Theorem 4 (Iterative Computation of Possible Minimum Winning Coalitions)**.** *Let μ be a PCG and $C \in \mathfrak{U}(\mu)$ a minimum winning coalition from the unlabeled set of μ. The number of possible minimum winning coalitions of size k of the PCG $\mu \cup \{C\}$ can be obtain from $\mathfrak{P}(\mu, k)$ by:*

$$\begin{aligned}\mathfrak{P}(\mu \cup \{C\}, k) = &\,\mathfrak{P}(\mu, k) - \mathbb{1}_{|C|=k}\\ &+ \mathbb{1}_{|C|<k} \cdot \sum_{\mathbb{S} \subseteq \mu^{k-}} (-1)^{|\mathbb{S}|+1} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}\\ &+ \mathbb{1}_{|C|>k} \cdot \sum_{\mathbb{S} \subseteq \mu^{k+}} (-1)^{|\mathbb{S}|+1} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cap C\right|}{k} \qquad (10)\end{aligned}$$

Sketch of Proof. Three cases are derived, based on if $|C|$ is strictly greater, strictly smaller and equal to k. When writing Equation (9) for $\mu \cup \{C\}$, only the part of the Equation corresponding to the case changes from $\mathfrak{P}(\mu, k)$. In that part, the sum over the subsets of $\mu \cup \{C\}$ can be split in two: one where C is present and one where it isn't, which gives us back the expression of $\mathfrak{P}(\mu, k)$ plus a sum. This sum will be over C itself and all unions of C with the subsets of μ, which can be regrouped in one term by dropping $\mathbb{S} \neq \emptyset$ in the sum definition. The full proof is given in Appendix B ☐

Note that these results hold for any voting game, not only μ^i and μ^c. Enumerating all PIXs or CfXs is then costly as their maximum theoretical number is combinatorial (Theorem 1).

We will see, in Section 5, how to compute the Shapley value of a PCG, *i.e.* knowing its value during a step of the enumeration of explanations, and in Section 6 how to compute bounds for the Shapley value from $\mathfrak{C}(\mu)$.

5 COMPUTING THE SHAPLEY VALUE OF A COMPACT GAME

In this Section we will see how the Shapley value of a feature for the decision $M(\mathbf{x})$ can be computed from the definition of a compact game, i.e. the set of minimum winning coalitions. As the Shapley value is computed on a game where the values of all coalitions are set, we will treat each Partial Compact Game as the Compact Game that is perceived, *i.e.* the Shapley value is computed like the information that the enumeration of PIXs or CfXs is ended has been received. In this case, all the coalitions in the unlabeled set of the PCG are considered to have a value of 0. If a new minimum winning coalition is added, we will have to correct the Shapley value by the impact of flipping the game's value of this coalition (and probably more supersets) to 1.

Let's begin by identifying the easy case of a Compact Game composed by a single minimum winning coalition. Then, we generalize the computation to any Compact Game through decomposition and finally derive an iterative computation, allowing to update the Shapley value upon the arrival of a new minimum winning coalition.

5.1 SINGLE MINIMUM WINNING COALITION

Let's start with the computation of the most simple compact games, featuring a single minimum winning coalition, that we will call a *simple compact game*.

Proposition 2 (Shapley Value of Simple Compact Games)**.** *The Shapley value for a feature i on a simple CG $\mu = \{\mathcal{S}\}$ is computed by:*

$$\phi_i(\nu^\mu) = \begin{cases} \frac{1}{|\mathcal{S}|} & \textit{if } i \in \mathcal{S}, \\ 0 & \textit{otherwise.} \end{cases} \tag{11}$$

Proof. By definition of ν^μ, $C \in 2^{\mathcal{N}}$ $\nu^\mu(C) = 1 \Leftrightarrow \mathcal{S} \subseteq C$. It follows that in particular $\forall C \in 2^{\mathcal{N}}$ $\mathcal{S} \subseteq C \Rightarrow \forall j \in \mathcal{S}$ $\nu^\mu(C \setminus \{j\}) = 0$, *i.e.* only features in $\mathcal{S}$ are pivotal. It follows that by the null player axiom, $\forall i \in \mathcal{N}, i \notin \mathcal{S} \Rightarrow \phi_i(\nu^\mu) = 0$, and by the symmetry axiom, all features in S must have the same Shapley value, *i.e.* $\forall i \in \mathcal{N}, i \in \mathcal{S} \Rightarrow \phi_i(\nu^\mu) = 1/|\mathcal{S}|$. A proof by calculation is given in Appendix C. □

5.2 DECOMPOSING COMPACT GAMES

Let's first consider the case where $\mu = \{\mathcal{S}_1, \mathcal{S}_2\}$ before generalizing to any CG. By definition:

$$\nu^\mu(C) = \begin{cases} 1 & \text{if } \mathcal{S}_1 \subseteq C \text{ or (non exclusive) } \mathcal{S}_2 \subseteq C, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

If we try to decompose μ in two subgames, $\mu_1 = \{\mathcal{S}_1\}$ and $\mu_2 = \{\mathcal{S}_2\}$, the sum of their value functions will not exactly coincide with ν^μ because:

$$\nu^{\mu_1}(C) + \nu^{\mu_2}(C) = \begin{cases} 2 & \text{if } \mathcal{S}_1 \subseteq C \text{ and } \mathcal{S}_2 \subseteq C, \\ 1 & \text{if } \mathcal{S}_1 \subseteq C \text{ and } \mathcal{S}_2 \not\subseteq C, \\ 1 & \text{if } \mathcal{S}_1 \not\subseteq C \text{ and } \mathcal{S}_2 \subseteq C, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

i.e. coalitions C s.t. $\mathcal{S}_1 \cup \mathcal{S}_2 \subseteq C$ are counted twice. It is important to note that it corresponds to the simple CG $\mu_{1,2} = \{\mathcal{S}_1 \cup \mathcal{S}_2\}$. The decomposition can then be generalized with Theorem 5.

Theorem 5 (Game Decomposition)**.** *The prediction game ν^μ, based on the CG μ, can be decomposed into a linear sum of simple CGs:*

$$\nu^\mu = \sum_{\substack{\mathbb{S} \subseteq \mu \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|+1} \cdot \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}} \tag{14}$$

Sketch of Proof. Consider all coalitions C which are superset of exactly $l \in \{1, \ldots, |\mu|\}$ coalitions in μ (if C is not a superset of any coalition in μ, its value in any game is trivially 0), and denote this set $\mu^l \subseteq \mu$. By definition of inclusion $\forall \mathbb{S}^l \subseteq \mu^l$, $C \supseteq \bigcup_{\mathcal{S} \in \mathbb{S}^l} \mathcal{S}$, so C has a value of 1 in each associated simple game thanks to monotonicity, and is not a superset of any union of minimum winning coalitions with $\mathbb{S} \subseteq \mu \setminus \mu^l$, with a value in the associated game of 0. When counting and summing the right hand side of Equation (14), 1 is obtained, *i.e.* the value of $\nu^\mu(C)$. The full proof is given in Appendix D. □

This decomposition could be interpreted as a special case of k-weighted voting game Brandt et al. [2016]. Now that a game can be decomposed into a linear form of its minimum winning coalition, it is easy to find the formula of the Shapley value of a feature:

Theorem 6 (Shapley Value of a Compact Game)**.** *The Shapley value of feature i in CG μ is:*

$$\phi_i(\nu^\mu) = \sum_{\substack{\mathbb{S} \subseteq \mu \\ \mathbb{S} \neq \emptyset}} \frac{(-1)^{|\mathbb{S}|+1}}{\left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}} \tag{15}$$

Proof. The result is obtained by applying the additivity axiom on Equation (14) and replacing the Shapley value of the simple compact games by their value in Equation (11). □

The decomposition of the computation of the Shapley value as a sum of smaller more specific games reminds the work of Conitzer and Sandholm [2004]. In our setting, the computation of the Shapley value is more simplified (linear in the number of players), trading the complexity in the size of the sum. It can also be used to write an MC-net Ieong and Shoham [2005] with rules $\forall \mathcal{S} \subseteq \mu$ s.t. $\mathcal{S} \neq \emptyset, \{\wedge_{i \in \mathcal{S}} i\} \rightarrow (-1)^{|\mathcal{S}|+1}$.This decomposition allows us to compute the entire game by only considering a strict subset of $2^{\mathcal{N}}$, composed by the unions of elements from μ.

5.3 DEFINING ITERATIVE COMPUTATION

Because any game can be decomposed into a linear form of its minimum winning coalition, the difference in the expression of the Shapley value of two compact games which only differ by one minimum winning coalition could be turned into an iterative computation.

Theorem 7 (Iterative Computation of the Shapley Value)**.** *Let μ be a PCG and $C \in \mathfrak{U}(\mu)$ a minimum winning coalition from the unlabeled set of μ. The Shapley value of feature i in CG $\mu \cup \{C\}$ can be obtained from $\phi_i(\nu^{\mu})$ by:*

$$\begin{aligned} \phi_i(\nu^{\mu \cup \{C\}}) &= \phi_i(\nu^{\mu}) \\ &+ \sum_{\mathbb{S} \subseteq \mu} \frac{(-1)^{|\mathbb{S}|}}{\left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C} \end{aligned} \tag{16}$$

Sketch of Proof. We apply Theorem 6 to get the expression of $\phi_i(\nu^{\mu \cup \{C\}})$ and a decomposition similar to the proof of Theorem 4 between the coalitions of the powerset of $\mu \cup \{C\}$ with and without C. The full proof is given in Appendix E. □

Even though feature $i \notin C$, the fact that it belongs to previous minimum winning coalitions of μ will update its Shapley value. It is then not necessary to compute all unions of any number of coalitions of μ, only those which contains i. Computations might also be improved by the fact that the unions will converge towards the big coalition $\mathcal{N}$, which can prevent unnecessary computations.

Consequently, during the enumeration of the PIXs or CfXs, the value associated to a feature might highly vary, and even ordering the contributions might require the elicitation of a large number of explanations, which is to be avoided. Next section aims at solving this issue by computing lower and upper bounds on the Shapley value of each feature, which could be used to build an interval order between features.

6 ESTIMATING BOUNDS OF THE SHAPLEY VALUE

In this section, we aim at finding bounds on the features Shapley value w.r.t the possible completions of the Partial Compact Game. We split this section in two parts. In the first we find the tightest bounds by identifying the best and worst case Compact Games in $\mathfrak{C}(\mu)$. In the second, as computing the Shapley value on a large Compact Game is difficult, we build lower and upper bounds that are easily computable, yet less tight w.r.t the possible values the Shapley value can effectively take.

6.1 TIGHTEST BOUNDS

Let's define the following two CGs $\overline{\mu_i}$ and $\underline{\mu_i}$.

Definition 8 (Optimistic Completion of a Partial Compact Game)**.** *Let μ be a PCG, and $i \in \mathcal{N}$ be a feature. Let $\overline{\mu_i} \in \mathfrak{C}(\mu)$ be the completion of μ defined by:*

$$\overline{\mu_i} = \mu \cup \{C \in \mathfrak{U}(\mu) \mid i \in C,\ \forall j \in C,\ C \setminus \{j\} \notin \mathfrak{U}(\mu)\} \tag{17}$$

Definition 9 (Pessimistic Completion of a Partial Compact Game)**.** *Let μ be a PCG, and $i \in \mathcal{N}$ be a feature. Let $\underline{\mu_i} \in \mathfrak{C}(\mu)$ be the completion of μ defined by:*

$$\underline{\mu_i} = \mu \cup \{C \in \mathfrak{U}(\mu) \mid i \notin C, \forall C' \in \mathfrak{U}(\mu),\ C' \not\subset C\} \tag{18}$$

These two CG corresponds to the best and worst possible completions of μ w.r.t the Shapley value of the feature i.

Theorem 8 (Tightest Upper Bound)**.** *Let μ be a PCG, and $i \in \mathcal{N}$ be a feature. The best Shapley value for feature i, denoted $\overline{\phi_i}$, is defined by:*

$$\overline{\phi_i} = \max_{\mu' \in \mathfrak{C}(\mu)} \phi_i(\nu^{\mu'}) = \phi_i(\nu^{\overline{\mu_i}}) \tag{19}$$

Sketch of Proof. Because adding $C \in \mathfrak{U}(\mu)$ s.t. $i \in C$ strictly improves the Shapley value of feature i, and that adding C instead of C', s.t. $C \subset C'$ increases the Shapley value of i strictly more, the best CG in $\mathfrak{C}(\mu)$ is μ increased by the antichain of coalitions containing i from $\mathfrak{U}(\mu)$ which has the strongest increase in Shapley Value, *i.e.* $\overline{\mu_i}$. The full proof is given in Appendix F. □

Theorem 9 (Tightest Lower Bound)**.** *Let μ be a PCG, and $i \in \mathcal{N}$ be a feature. The lowest Shapley value for feature i, denoted $\underline{\phi_i}$, is defined by:*

$$\underline{\phi_i} = \min_{\mu' \in \mathfrak{C}(\mu)} \phi_i(\nu^{\mu'}) = \phi_i(\nu^{\underline{\mu_i}}) \tag{20}$$

Sketch of Proof. The worst game in $\mathfrak{C}(\mu)$ w.r.t feature i must only contain the coalitions $\mathcal{S} \in 2^{\mathcal{N}}$ where i is pivotal in ν^{μ} and for which it is not possible to choose $C \in \mathfrak{U}(\mu)$ s.t. $C \subseteq \mathcal{S} \setminus \{i\}$. This is the case for $\underline{\mu_i}$ which is μ increased by the antichain of coalitions not containing i from $\mathfrak{U}(\mu)$ which maximizes the score of the other features. The full proof is given in Appendix G. □

The minimum of Shapley value for feature i can be found on other compact games definition, for instance by computing all the coalitions for which i is pivotal in μ and find (if it exists) explicitly the corresponding coalition of $\mathfrak{U}(\mu)$ which is minimal w.r.t inclusion.

Regardless, the iterative update of such games following the enumeration of PIXs or CfXs might prove difficult to compute, especially without an explicit form of $\mathfrak{U}(\mu)$.

6.2 COMPUTABLE LOWER AND UPPER BOUNDS

The tightest bound computation requires to compute the Shapley value on Compact Games that are large and difficult to build, even way larger than the PCG μ at hand. Lower and upper approximation of these values that are easy to compute can however be designed.

Definition 10 (Lower Bound Approximation)**.** *Let $\underline{\underline{\phi_i}}$ be defined by:*

$$\underline{\underline{\phi_i}}(\nu^\mu) = \sum_{\substack{\mathcal{S} \in \mu \\ i \in \mathcal{S}}} c(\mathcal{S} \setminus \{i\}) \tag{21}$$

$\underline{\underline{\phi_i}}(\nu^\mu)$ is then the score (using Equation (1)) obtained by i by being a pivotal feature for the minimum winning coalitions it belongs to. This score corresponds to what cannot be removed from the Shapley value of feature i if new coalitions from the unlabeled set $\mathfrak{U}(\mu)$ are added to μ. It is then a lower bound of the true minimum $\underline{\phi_i}(\nu^\mu)$ value.

It is then easy to deduce a corresponding upper bound.

Definition 11 (Upper Bound Approximation)**.** *Let $\overline{\overline{\phi_i}}$ be defined by:*

$$\overline{\overline{\phi_i}}(\nu^\mu) = 1 - \sum_{\substack{j \in \mathcal{N} \\ j \neq i}} \underline{\underline{\phi_j}}(\nu^\mu) \tag{22}$$

$\overline{\overline{\phi_i}}(\nu^\mu)$ corresponds then to the remaining total value to distribute after setting the Shapley value of the other features to their lower approximation. It is then an upper bound of what i could receive.

It is then trivial to see that the iterative update has a linear cost and that the lower bound approximation is increasing and the upper bound approximation is decreasing, yet most likely not very tightly w.r.t their true values.

7 CONCLUSION AND LIMITATIONS

We have presented two collaborative games to formally use the Shapley value as a way to quantify the impact of a feature in the result of a classification decision made by a model M on an input vector $\mathbf{x}$. One game is based on the Prime Implicant explanations of this decision, and the second one is based on the subset minimal Counterfactual explanations. Doing so allows for the Shapley value to measure the impact of a feature $i \in \mathcal{N}$ being observed or changed from its value $\mathbf{x}_i$ in the observation, which gives us a measure of the local importance of the feature.

We have shown that our two games are normalized and monotone, which allows us to connect with the social choice literature. These well studied "voting games" help us to give an interpretation to the Shapley value of a feature: it is the probability of the feature to be decisive for the prediction. Although we correct some problems raised by SHAP-based methods Fryer et al. [2021], Janzing et al. [2020], e.g. w.r.t sampling, some others are inherent to the use of the Shapley value in explainability: the paradox of the increase/decrease of power by splitting identities, well known in social choice Brandt et al. [2016]. In voting games if a voter is duplicated by dividing its weight in two entities, there is an increase in the Shapley value shared by the two and a decrease in the Shapley value of the other voters. In our explainability context, correlation between features will have this effect, which makes the use of the Shapley value to reflect the real contribution of a feature less reliable. Also the Shapley value will not tell much about the "true" strength of a feature. For example take two games, one where each feature is sufficient to take a decision and a second where we need two features to take the decision. In both cases, the features will receive the same Shapley value of $1/n$.

Building on the Prime Implicant and Counterfactual explanations enumeration methods Marques-Silva et al. [2020], de Colnet and Marquis [2022], Audemard et al. [2024] we presented (i) how to compute the maximum number of remaining explanations in the enumeration, (ii) given a step in the enumeration how to compute the precise Shapley value, (iii) bounds on this value, and (iv) iterative computation to update the values during the enumeration. We proved that all these could be obtained based on the list of Prime Implicant/Counterfactual explanations only.

This work presents the theoretical analysis of whether it was possible to avoid sampling and relearning of the model to obtain a rigorous cooperative game through a link with Prime Implicant and Counterfactual explanations. The answer is yes. Future work will then be focused on the comparison with other feature attribution methods for explainability Lundberg and Lee [2017], Létoffé et al. [2025], Kolpaczki et al. [2024]. We already paved the way, by finding bounds on the Shapley value, to answering the question "How long should we enumerate explanations to have a real idea of the features importance ?". Such comparative experiments shall also be used to see how fast we can isolate the top$-k$ features, whether other methods find at least the same order between features and how fast we can compute in practice. We could also imagine more efficient computations if we have structural properties on the enumeration of explanations (e.g. by size, by orders in the features etc.), or through approximation Kempinski and Kachman [2025].

References

Gilles Audemard, Frédéric Koriche, and Pierre Marquis. On Tractable XAI Queries based on Compiled Representations. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, pages 838–849, 9 2020.

Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the Computational Intelligibility of Boolean Classifiers. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*, pages 74–86, 11 2021.

Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. PyXAI: An XAI library for tree-based models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea*, pages 8601–8605, 2024.

JM Bilbao, JR Fernandez, A Jiménez Losada, and JJ Lopez. Generating functions for computing power indices efficiently. *Top*, 8(2):191–213, 2000.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

Martin Cerný and Michel Grabisch. Incomplete cooperative games with player-centered information. *Discret. Appl. Math.*, 346:62–79, 2024.

Vincent Conitzer and Tuomas Sandholm. Computing Shapley values, manipulating value division schemes, and checking core membership in multi-issue domains. In Deborah L. McGuinness and George Ferguson, editors, *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*, pages 219–225. AAAI Press / The MIT Press, 2004.

Adnan Darwiche and Pierre Marquis. On Quantifying Literals in Boolean Logic and Its Applications to Explainable AI. *Journal of Artificial Intelligence Research*, 72:285–328, 2021.

Alexis de Colnet and Pierre Marquis. On the complexity of enumerating prime implicants from decision-DNNF circuits. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria*, pages 2583–2590, 2022.

Alexis de Colnet and Pierre Marquis. On translations between ML models for XAI purposes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, Macao, SAR, China*, pages 3158–3166, 2023.

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

S. Shaheen Fatima, Michael J. Wooldridge, and Nicholas R. Jennings. A randomized method for the shapley value for the voting game. In Edmund H. Durfee, Makoto Yokoo, Michael N. Huhns, and Onn Shehory, editors, *6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007), Honolulu, Hawaii, USA, May 14-18, 2007*, page 157. IFAAMAS, 2007.

Daniel Vidali Fryer, Inga Strümke, and Hien D. Nguyen. Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360, 2021.

Ian Gemp, Marc Lanctot, Luke Marris, Yiran Mao, Edgar A. Duéñez-Guzmán, Sarah Perrin, András György, Romuald Elie, Georgios Piliouras, Michael Kaisers, Daniel Hennes, Kalesha Bullard, Kate Larson, and Yoram Bachrach. Approximating the core via iterative coalition sampling. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 669–678. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024.

Xuanxiang Huang and Joao Marques-Silva. On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning*, 171:109112, 2024. Synergies between Machine Learning and Reasoning.

Samuel Ieong and Yoav Shoham. Marginal contribution nets: a compact representation scheme for coalitional games. In John Riedl, Michael J. Kearns, and Michael K. Reiter, editors, *Proceedings 6th ACM Conference on Electronic Commerce (EC-2005), Vancouver, BC, Canada, June 5-8, 2005*, pages 193–202. ACM, 2005.

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva. From contrastive to abductive explanations and back again. In Matteo Baldoni and Stefania Bandini, editors, *AIxIA 2020 - Advances in Artificial Intelligence - XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, 2020, Revised Selected Papers*, volume 12414 of *Lecture Notes in Computer Science*, pages 335–355. Springer, 2020.

Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. Feature relevance quantification in explainable AI: A causal problem. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 2020.

Benjamin Kempinski and Tal Kachman. Influencenet: AI models for Banzhaf and Shapley value prediction. *CoRR*, abs/2503.08381, 2025.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the Shapley value without marginal contributions. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, Vancouver, Canada*, pages 13246–13255. AAAI Press, 2024.

Olivier Létoffé, Xuanxiang Huang, and João Marques-Silva. Towards trustable SHAP scores. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, Philadelphia*, pages 18198–18208. AAAI Press, 2025.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *NeurIPS 2020, virtual*, volume 33, pages 20590–20600, 2020.

Satoshi Masuya. An approximated Shapley value for partially defined cooperative games. In Jaroslaw Wątróbski, Wojciech Salabun, Carlos Toro, Cecilia Zanni-Merk, Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES-2021, Virtual Event / Szczecin, Poland, 8-10 September 2021*, Procedia Computer Science, pages 100–108. Elsevier, 2021.

Satoshi Masuya and Masahiro Inuiguchi. A fundamental study for partially defined cooperative games. *Fuzzy Optim. Decis. Mak.*, 15(3):281–306, 2016.

Tomasz P. Michalak, Aadithya V. Karthik, Piotr L. Szczepanski, Balaraman Ravindran, and Nicholas R. Jennings. Efficient computation of the Shapley value for game-theoretic network centrality. *J. Artif. Intell. Res.*, 46: 607–650, 2013.

Anatol Rapoport. *N-person game theory: Concepts and applications*. Dover Publications, Mineola, NY, 2001.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Oliver Kiss, Sebastian Nilsson, and Rik Sarkar. The Shapley value in machine learning. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria*, pages 5572–5579, 2022.

L. S. Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, 48(3):787–792, 1954.

Andy Shih, Arthur Choi, and Adnan Darwiche. Compiling Bayesian network classifiers into decision graphs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii*, pages 7966–7974. AAAI Press, 2019.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery.

Sofiane Touati, Mohammed Said Radjef, and Lakhdar Sais. A bayesian Monte Carlo method for computing the Shapley value: Application to weighted voting and bin packing games. *Comput. Oper. Res.*, 125:105094, 2021.

Filip Úradník, David Sychrovský, Jakub Cerný, and Martin Cerný. Reducing optimism bias in incomplete cooperative games. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 1847–1855. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024.

Stephen J Willson. A value for partially defined cooperative games. *International Journal of Game Theory*, 21(4): 371–384, 1993.

Building a Bridge Between the Shapley Value, Prime Implicants and Counterfactual Explanations: a Theoretical Analysis (Supplementary Material)

Hénoïk Willot[1]

[1]Sorbonne Université, CNRS, LIP6 , Paris, France

A NON SUB AND SUPER MODULARITY OF THE GAMES

It is easy to show, using the definitions of PIXs and CfXs, that prediction games ν^i and ν^c do not verify (in the general case) sub or super modularity.

Proposition 3 (Submodularity)**.** *In the general case, neither ν^i or ν^c is* submodular*, i.e. it is false that:*

$$\forall C, D \subseteq \mathcal{N},\ \nu(C) + \nu(D) \geq \nu(C \cup D) + \nu(C \cap D).$$

Proof. Let $C \cup D \subseteq \mathcal{N}$, s.t. $C \neq D$, be a PIX :

- $\nu^i(C) = 0$ and $\nu^i(D) = 0$, yet
- $\nu^i(C \cup D) = 1$ and $\nu^i(C \cap D) = 0$.

The same property holds for CfXs and ν^c. □

Proposition 4 (Supermodularity)**.** *In the general case, neither ν^i or ν^c is* supermodular*, i.e. it is false that:*

$$\forall C, D \subseteq \mathcal{N},\ \nu(C) + \nu(D) \leq \nu(C \cup D) + \nu(C \cap D).$$

Proof. Let $C, D \subseteq \mathcal{N}$, s.t. $C \neq D$, be two PIXs:

- $\nu^i(C) = 1$ and $\nu^i(D) = 1$, yet
- $\nu^i(C \cup D) = 1$ and $\nu^i(C \cap D) = 0$.

The same property holds for CfXs and ν^c. □

B PROOF OF THEOREM 4

Proof. The aim is to prove that given a PCG μ and $C \in \mathfrak{U}(\mu)$ a minimum winning coalition from the unlabeled set of μ, the number of possible minimum winning coalitions of size k of the PCG $\mu \cup \{C\}$ can be obtain from $\mathfrak{P}(\mu, k)$ by:

$$\mathfrak{P}(\mu \cup \{C\}, k) = \mathfrak{P}(\mu, k) - \mathbb{1}_{|C|=k} + \mathbb{1}_{|C|<k} \cdot \sum_{\mathbb{S} \subseteq \mu^{k-}} (-1)^{|\mathbb{S}|+1} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}$$
$$+ \mathbb{1}_{|C|>k} \cdot \sum_{\mathbb{S} \subseteq \mu^{k+}} (-1)^{|\mathbb{S}|+1} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cap C\right|}{k} \tag{23}$$

We recall μ^{k+}, μ^{k-} and μ^{k}, the subsets of PCG μ, forming a partition, such that they contains the coalitions of size respectively strictly greater, strictly smaller and equal to k. The positioning of C w.r.t these sets will have an impact. Let's investigate the three cases:

1. $|C| < k$
2. $|C| = k$
3. $|C| > k$

The easiest case is case 2. Let $\mu' = \mu \cup \{C\}$. It follows from Theorem 2:

$$\begin{aligned}
\mathfrak{P}(\mu', k) &= \binom{n}{k} + \sum_{\substack{\mathbb{S} \subseteq \mu'^{k-} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} - |\mu'^{k}| + \sum_{\substack{\mathbb{S} \subseteq \mu'^{k+} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k} \\
&= \binom{n}{k} + \sum_{\substack{\mathbb{S} \subseteq \mu^{k-} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} - |\mu^{k}| - 1 + \sum_{\substack{\mathbb{S} \subseteq \mu^{k+} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k} \\
&= \mathfrak{P}(\mu, k) - 1
\end{aligned}$$

i.e. the impact of adding C to the minimum winning coalition set is only to remove it from the number of available choices of coalitions of size $k = |C|$.

For the other two cases, $|C| \neq k$ so the impact of C on coalitions of size exactly k is to be considered. First, C could not have an impact on coalitions $\mathcal{S} \in \mu$ of size $|\mathcal{S}| = k$, as it would mean that $\mathcal{S} \subseteq C$, *i.e.* C is not a minimum winning coalition.

Let's first consider case 1 where $|C| < k$. The impact of C on coalitions of size exactly k is to remove all coalitions that are superset of C, which can be counted easily as we only have to choose $k - |C|$ features, to build a coalition of size k which is a superset of C, from the remaining $n - |C|$ features. However some of these coalitions of size k might already have been counted out by some other minimum coalition of μ. Not from μ^{k}, as it would contradict the fact that C is a minimum winning coalition. It also cannot be with some $\mathcal{S} \in \mu^{k+}$, as it would require to find a coalition D such that $D \subseteq \mathcal{S}$ and $C \subseteq D$, yet it is impossible to have $C \subseteq \mathcal{S}$ as one of the two would not be a minimum winning coalition. Interaction can only come from minimum winning coalitions of μ^{k-}.

It then follows from Theorem 2:

$$\begin{aligned}
\mathfrak{P}(\mu', k) &= \binom{n}{k} + \sum_{\substack{\mathbb{S} \subseteq \mu'^{k-} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} - |\mu'^{k}| + \sum_{\substack{\mathbb{S} \subseteq \mu'^{k+} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k} \\
&= \binom{n}{k} - |\mu^{k}| + \sum_{\substack{\mathbb{S} \subseteq \mu^{k+} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k} + \sum_{\substack{\mathbb{S} \subseteq \mu^{k-} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} \\
&\quad + \sum_{\mathbb{S} \subseteq \mu^{k-}} (-1)^{|\mathbb{S}|+1} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|} \\
&= \mathfrak{P}(\mu, k) + \sum_{\mathbb{S} \subseteq \mu^{k-}} (-1)^{|\mathbb{S}|+1} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|}
\end{aligned}$$

by splitting the powerset of $\mu \cup \{C\}$ enumeration between subsets containing C and the rest.

With exactly the same reasoning about the third case 3, the interaction can only be found with coalitions in μ^{k+}, and separation of the powerset of $\mu \cup \{C\}$ enumeration in the sum. We obtain:

$$\begin{aligned}\mathfrak{P}(\mu', k) &= \binom{n}{k} + \sum_{\substack{\mathbb{S} \subseteq \mu'^{k-} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{n - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k - \left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} - |\mu'^{k}| + \sum_{\substack{\mathbb{S} \subseteq \mu'^{k+} \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|}{k} \\ &= \mathfrak{P}(\mu, k) + \sum_{\mathbb{S} \subseteq \mu^{k+}} (-1)^{|\mathbb{S}|+1} \binom{\left|\bigcap_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cap C\right|}{k}\end{aligned}$$

Using the indicator functions to put together the three cases into a single equation, we obtain Equation (23)

□

C PROOF OF PROPOSITION 2

Proof. The proof given in the core of the paper uses the axiomatic properties of the Shapley value and the monotonicity of the game. The aim is to prove, by calculation, that the Shapley value for a feature i on a simple CG $\mu = \{\mathcal{S}\}$ is computed by:

$$\phi_i(\nu^{\mu}) = \begin{cases} \frac{1}{|\mathcal{S}|} & \text{if } i \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

Recalling that for $\mu = \{\mathcal{S}\}$, $\nu^{\mu}(C) = 1 \Leftrightarrow \mathcal{S} \subseteq C$, it is naturally that $\nu^{\mu}(C) = 0 \Leftrightarrow \mathcal{S} \nsubseteq C$. For the Shapley value, the only coalitions to consider are those for which adding i triggers the change of value from 0 to 1, *i.e.* only coalitions $C = D \setminus \{i\}, \forall D \supseteq \mathcal{S}$ are considered. As the contribution to the Shapley value of a coalition only depends on its size (as the marginal contribution to the game is 1), the Shapley value can be directly rewritten as a count of such coalitions.

$$\begin{aligned}\phi_i(\mu) &= \sum_{C \subseteq \mathcal{N} \setminus \{i\}} c(C) \cdot (\nu^{\mu}(C \cup \{i\}) - \nu^{\mu}(C)) \\ &= \sum_{k=|\mathcal{S}|}^{n} \frac{(k-1)! \cdot (n-(k-1)-1)!}{n!} \cdot \binom{n-|\mathcal{S}|}{k-|\mathcal{S}|} \\ &= \sum_{k=|\mathcal{S}|}^{n} \frac{(n-|\mathcal{S}|)! \cdot (k-1)! \cdot (n-k)!}{(n-k)! \cdot (k-|\mathcal{S}|)! \cdot n!} \\ &= \sum_{k=|\mathcal{S}|}^{n} \frac{(k-1)! \cdot (n-|\mathcal{S}|)! \cdot (|\mathcal{S}|-1)!}{(k-|\mathcal{S}|)! \cdot (|\mathcal{S}|-1)! \cdot n \cdot (n-1)!} \\ &= \sum_{k=|\mathcal{S}|}^{n} \frac{1}{n} \cdot \frac{\binom{k-1}{|\mathcal{S}|-1}}{\binom{n-1}{|\mathcal{S}|-1}} \\ &= \frac{1}{n} \cdot \frac{1}{\binom{n-1}{|\mathcal{S}|-1}} \cdot \sum_{k=|\mathcal{S}|}^{n} \binom{k-1}{|\mathcal{S}|-1} \\ &= \frac{1}{n} \cdot \frac{\binom{n}{|\mathcal{S}|}}{\binom{n-1}{|\mathcal{S}|-1}} \quad \text{[Hockey stick identity]} \\ &= \frac{1}{n} \cdot \frac{n}{|\mathcal{S}|} \\ &= \frac{1}{|\mathcal{S}|}\end{aligned}$$

□

D PROOF OF THEOREM 5

Proof. The aim is to prove that the prediction game ν^{μ}, based on the CG μ, can be decomposed into a linear sum of simple compact games:

$$\nu^{\mu} = \sum_{\substack{\mathbb{S} \subseteq \mu \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|+1} \cdot \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}} \tag{25}$$

Let's consider $C \subseteq \mathcal{N}$ s.t. $\forall \mathcal{S} \in \mu,\ C \not\supseteq \mathcal{S}$. Its value is then trivially 0 in ν^{μ} and in any $\nu^{\{\mathcal{S}\}}, \mathcal{S} \in \mu$, so the decomposition is true.

Let's now consider $C \subseteq \mathcal{N}$ and $\mu' \subseteq \mu$ s.t. $\mu' = \{\mathcal{S} \in \mu \mid C \supseteq \mathcal{S}\} \neq \emptyset$, so μ' contains all the minimum winning coalitions of which C is a superset, *i.e.* where it is counted more than once.

By definition of the inclusion, of μ' and by the monotonicity of games we have:

- $\forall \mathbb{S} \subseteq \mu',\ C \supseteq \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \Rightarrow \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}}(C) = 1$
- $\forall \mathbb{S} \subseteq \mu,\ C \not\supseteq \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \Rightarrow \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}}(C) = 0$
- $\forall \mathbb{S} \subseteq \mu \setminus \mu'\ \forall \mathbb{S}' \subseteq \mu',\ C \not\supseteq \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup \bigcup_{\mathcal{S}' \in \mathbb{S}'} \mathcal{S}' \Rightarrow \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup \bigcup_{\mathcal{S}' \in \mathbb{S}'} \mathcal{S}'\}}(C) = 0$

The 1's can then be counted by size of subsets of μ. There are two cases, one considering a union of less minimum winning coalitions than the number that belong to μ':

$$\begin{aligned}
&\forall k \in \{1, \ldots, |\mu'|\},\ \sum_{\substack{\mathbb{S} \subseteq \mu \\ |\mathbb{S}|=k}} \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}}(C) = \sum_{\substack{\mathbb{S}' \subseteq \mu' \\ |\mathbb{S}'|=k}} \nu^{\{\bigcup_{\mathcal{S}' \in \mathbb{S}'} \mathcal{S}'\}}(C) = \binom{|\mu'|}{k} \\
\Rightarrow &\forall k \in \{1, \ldots, |\mu'|\},\ \sum_{\substack{\mathbb{S}' \subseteq \mu' \\ |\mathbb{S}'|=k}} (-1)^{k+1} \nu^{\{\bigcup_{\mathcal{S}' \in \mathbb{S}'} \mathcal{S}'\}}(C) = (-1)^{k+1} \binom{|\mu'|}{k}
\end{aligned}$$

and one considering a union of $k > |\mu'|$ minimum winning coalitions, where at least one minimum winning coalition in $\mu \setminus \mu'$ has to be picked, hence the game value for C is 0.

By summing over all values of k, we obtain:

$$\begin{aligned}
\sum_{k=1}^{|\mu|} \sum_{\substack{\mathbb{S} \subseteq \mu \\ |\mathbb{S}|=k}} (-1)^{k+1} \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}}(C) &= \sum_{k=1}^{|\mu'|} (-1)^{k+1} \binom{|\mu'|}{k} + \sum_{|\mu'|+1}^{n} 0 \\
\Rightarrow \sum_{\substack{\mathbb{S} \subseteq \mu \\ \mathbb{S} \neq \emptyset}} (-1)^{|\mathbb{S}|+1} \cdot \nu^{\{\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\}}(C) &= -(-1)^{1} \binom{|\mu'|}{0} + \sum_{k=0}^{|\mu'|} (-1)^{k+1} \binom{|\mu'|}{k} \\
&= 1 + (1-1)^{|\mu'|} \\
&= 1
\end{aligned}$$

which concludes the proof. □

E PROOF OF THEOREM 7

Proof. The aim is to prove that, given a PCG μ and $C \in \mathfrak{U}(\mu)$ a minimum winning coalition from the unlabeled set of μ, the Shapley value of feature i in CG $\mu \cup \{C\}$ can be obtain from $\phi_i(\nu^{\mu})$ by:

$$\phi_i(\nu^{\mu \cup \{C\}}) = \phi_i(\nu^{\mu}) + \sum_{\mathbb{S} \subseteq \mu} \frac{(-1)^{|\mathbb{S}|}}{\left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C} \tag{26}$$

Let $\mu' = \mu \cup C$. By applying Theorem 6 on μ' it follows that:

$$\phi_i(\nu^{\mu'}) = \sum_{\substack{\mathbb{S}' \subseteq \mu' \\ \mathbb{S}' \neq \emptyset}} \frac{(-1)^{|\mathbb{S}'|+1}}{\left|\bigcup_{\mathcal{S}' \in \mathbb{S}'} \mathcal{S}'\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S}' \in \mathbb{S}'} \mathcal{S}'}$$

This sum can be decomposed in two components, one for all subsets of the powerset of μ' which don't contain C, *i.e.* subsets from μ without the empty set, and the second for all subsets containing C, *i.e.* the union between C and the subsets of μ. It is then possible to write:

$$\begin{aligned} \phi_i(\nu^{\mu'}) &= \sum_{\substack{\mathbb{S} \subseteq \mu \\ \mathbb{S} \neq \emptyset}} \frac{(-1)^{|\mathbb{S}|+1}}{\left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S}} + \sum_{\mathbb{S} \subseteq \mu} \frac{(-1)^{|\mathbb{S}|+2}}{\left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C} \\ &= \phi_i(\nu^{\mu}) + \sum_{\mathbb{S} \subseteq \mu} \frac{(-1)^{|\mathbb{S}|}}{\left|\bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C\right|} \cdot \mathbb{1}_{i \in \bigcup_{\mathcal{S} \in \mathbb{S}} \mathcal{S} \cup C} \end{aligned}$$

which concludes the proof. □

F PROOF OF THEOREM 8

Proof. The aim is to prove that given a PCG μ, a feature $i \in \mathcal{N}$ and $\overline{\mu_i} \in \mathfrak{C}(\mu)$ the completion of μ defined by:

$$\overline{\mu_i} = \mu \cup \{C \in \mathfrak{U}(\mu) \mid i \in C,\ \forall j \in C,\ C \setminus \{j\} \notin \mathfrak{U}(\mu)\} \tag{27}$$

then the best Shapley value for feature i, denoted $\overline{\phi_i}$, is defined by:

$$\overline{\phi_i} = \max_{\mu' \in \mathfrak{C}(\mu)} \phi_i(\nu^{\mu'}) = \phi_i(\nu^{\overline{\mu_i}}) \tag{28}$$

We begin the proof by proving the two following results :

Proposition 5 (Strict Increase in Shapley Value)**.** *Let μ be a CG and $C \in \mathfrak{U}(\mu)$ a coalition compatible with μ such that $i \in C$. We have:*

$$\phi_i(\nu^{\mu \cup \{C\}}) > \phi_i(\nu^{\mu}) \tag{29}$$

Proposition 6 (Smaller Coalition Means Higher Shapley Value)**.** *Let μ be a CG and $C \in \mathfrak{U}(\mu)$ a coalition compatible with μ such that $i \in C$. We have:*

$$\phi_i(\nu^{\mu \cup \{C\}}) > \phi_i(\nu^{\mu \cup \{C' \in \mathfrak{U}(\mu) \mid C \subset C'\}}) \tag{30}$$

i.e., adding C to μ has a strictly bigger impact than adding all supersets of C from $\mathfrak{U}(\mu)$.

We begin with Proposition 5. Let's first investigate ν^{μ}. As introduced in Section 3.2, the Shapley value of feature i in game ν^{μ} is defined by the coalitions for which i is pivotal, *i.e.* the coalitions $\mathcal{S} \subseteq \mathcal{N}$ s.t. $i \notin \mathcal{S}$, $\nu^{\mu}(\mathcal{S}) = 0$ and $\nu^{\mu}(\mathcal{S} \cup \{i\}) = 1$. In order to reduce the Shapley value of feature i by adding C to the CG, i must become not pivotal for at least one of these coalitions $\mathcal{S}$, *i.e.* verify $\nu^{\mu \cup \{C\}}(\mathcal{S}) = 1$.

Yet, the impact of adding C to the CG μ on the game values is to induce a flip of value from 0 to 1 for all coalitions C' such that $C' \supseteq C$ which had $\nu^{\mu}(C') = 0$. By definition $i \in C'$ and $i \notin \mathcal{S}$ so the flip is not occurring on any coalition $\mathcal{S}$ for which i is pivotal, so the Shapley value of i is non decreasing when adding coalition $C \in \mathfrak{U}(\mu)$ such that $i \in C$.

It can be seen that it is in fact strictly increasing because $\nu^{\mu}(C) = \nu^{\mu}(C \setminus \{i\}) = 0$ while $\nu^{\mu \cup \{C\}}(C) = 1$ and $\nu^{\mu \cup \{C\}}(C \setminus \{i\}) = 0$, *i.e.* i is pivotal in at least one supplementary coalition, so its Shapley value is increasing.

To sum up, for the Shapley value of i, it is always strictly valuable to add a coalition $C \in \mathfrak{U}(\mu)$ s.t. $i \in C$ to the CG μ. We will now show with Proposition 6 that, given the choice of the coalition $C \in \mathfrak{U}(\mu)$, s.t. $i \in C$, to add to μ, it is always strictly more profitable for the Shapley value of i to add the smallest C w.r.t inclusion in $\mathfrak{U}(\mu)$.

To prove Proposition 6, we begin by proving that given the choice between (1) adding $C \in \mathfrak{U}(\mu)$ such that $i \in C$ and (2) adding all coalitions $C' = C \cup \{j\}, \forall j \in \mathcal{N} \setminus C$, it is strictly more profitable to choose (1). This comparison corresponds to the best case where all $C' \in \mathfrak{U}(\mu)$ which is not necessarily the case as their value could have already been forced to 1 because of some $\mathcal{S} \in \mu$, *i.e.* the aim is to prove $\phi_i(\nu^{\mu \cup \{C\}}) > \phi_i(\nu^{\mu \cup \{C' \in 2^{\mathcal{N}} \mid C \subset C'\}})$.

Let $\mu_{(1)} = \mu \cup \{C\}$ and $\mu_{(2)} = \mu \cup \{C \cup \{j\} | \, j \in \mathcal{N} \setminus C\}$.

The associated games can be naturally defined as

- $\nu^{\mu^{(1)}} = \begin{cases} \nu^{\mu}(\mathcal{S}) & \text{if } C \not\subseteq \mathcal{S} \\ 1 & \text{if } C \subseteq \mathcal{S} \end{cases}$
- $\nu^{\mu^{(2)}} = \begin{cases} \nu^{\mu}(\mathcal{S}) & \text{if } \forall j \in \mathcal{N} \setminus C, \; C \cup \{j\} \not\subseteq \mathcal{S} \\ 1 & \text{if } \exists j \in \mathcal{N} \setminus C, \; C \cup \{j\} \subseteq \mathcal{S} \end{cases}$

It is easy to see that $\nu^{\mu^{(1)}}$ and $\nu^{\mu^{(2)}}$ only differ by $\nu^{\mu^{(1)}}(C) = 1$ and $\nu^{\mu^{(2)}}(C) = 0$. This means that in $\mu^{(1)}$ we do not have that $\forall j \in \mathcal{N} \setminus C, j$ is pivotal for C, as $\nu^{\mu^{(2)}}(C) = 0$ and $\nu^{\mu^{(2)}}(C \cup \{j\}) = 1$, which is irrelevant for i as $i \in C$. However, i is pivotal for C in $\mu^{(1)}$ as $\nu^{\mu^{(1)}}(C \setminus \{i\}) = 0$ and $\nu^{\mu^{(1)}}(C) = 1$ and not in $\mu^{(2)}$ as $\nu^{\mu^{(2)}}(C \setminus \{i\}) = 0$ and $\nu^{\mu^{(2)}}(C) = 0$. We obtain that i is pivotal exactly once more in $\mu^{(1)}$ than in $\mu^{(2)}$, so the Shapley value of i is not lower in $\mu^{(1)}$ than in $\mu^{(2)}$.

By induction, this holds when considering $\mu^{(1)}$ against strict supersets of all $C' = C \cup \{j\}, \forall j \in \mathcal{N} \setminus C$. We proved by leveraging the constraint on C' belonging to $\mathfrak{U}(\mu)$ the proposition is true, it is then true if the C' are more constrained, as in Equation (30). □

G PROOF OF THEOREM 9

Proof. The aim is to prove that given a PCG μ, a feature $i \in \mathcal{N}$ and $\underline{\mu_i} \in \mathfrak{C}(\mu)$ the completion of μ defined by:

$$\underline{\mu_i} = \mu \cup \{C \in \mathfrak{U}(\mu) \mid i \notin C, \forall C' \in \mathfrak{U}(\mu), \; C' \not\subset C\} \tag{31}$$

then the lowest Shapley value for feature i, denoted $\underline{\phi_i}$, is defined by:

$$\underline{\phi_i} = \min_{\mu' \in \mathfrak{C}(\mu)} \phi_i(\nu^{\mu'}) = \phi_i(\nu^{\underline{\mu_i}}) \tag{32}$$

The lowest Shapley value of feature i from a CG μ is obtained when the coalitions $\mathcal{S} \subseteq \mathcal{N}$ for which i is pivotal in μ are removed by the addition of some coalitions $C \in \mathfrak{U}(\mu)$.

As shown in Appendix F, adding to the CG μ a coalition $C \in \mathfrak{U}(\mu)$ s.t. $i \in C$ will strictly increase the Shapley value of i, so $\mu \cup \{C\}$ will not correspond to a minimum of the Shapley value of i as $\phi_i(\nu^{\mu}) < \phi_i(\nu^{\mu \cup \{C\}})$.

We can then focus only on $C \in \mathfrak{U}(\mu)$ s.t. $i \notin C$. As in our prediction games the total amount divided between the n features is 1, it can be said that:

$$\min_{\mu' \in \mathfrak{C}(\mu)} \phi_i(\nu^{\mu'}) = 1 - \max_{\mu' \in \mathfrak{C}(\mu)} \sum_{j \in \mathcal{N} \setminus \{j\}} \phi_j(\nu^{\mu'})$$

i.e. the sum of the Shapley value of the other features has to be maximized, while not adding coalitions $C \in \mathfrak{U}(\mu)$ s.t. $i \notin C$.

Similarly to the proof in Appendix F, it can be seen that:

Proposition 7 (Decrease in Shapley Value)**.** *Let μ be a CG and $C \in \mathfrak{U}(\mu)$ a coalition compatible with μ such that $i \notin C$. We have:*

$$\phi_i(\nu^{\mu \cup \{C\}}) \leq \phi_i(\nu^{\mu}) \tag{33}$$

Adding C to μ will set the value of all $C' \supset C$ to 1. So it is possible (not necessary, hence the greater or equal in Equation (33)) to find C' s.t. $\nu^{\mu}(C') = 0$ and $\nu^{\mu}(C' \cup \{i\}) = 1$ which became $\nu^{\mu \cup \{C\}}(C') = 1$ and $\nu^{\mu \cup \{C\}}(C' \cup \{i\}) = 1$, *i.e.* a coalition for which i was pivotal is removed, so its Shapley value decreases.

Together with Proposition 6 applied to features $j \in \mathcal{N} \setminus \{i\}$, the minimum of Shapley value for feature i is obtained when adding as many subset minimal coalitions $C \in \mathfrak{U}(\mu)$ such that $i \notin C$, *i.e.* for:

$$\underline{\mu_i} = \mu \cup \{C \in \mathfrak{U}(\mu) \mid i \notin C, \forall C' \in \mathfrak{U}(\mu), \; C' \not\subset C\}$$

□