Explaining Probabilistic Models with Distributional Values

Luca Franceschi [1] **Michele Donini** [1] **Cédric Archambeau** [2] **Matthias Seeger** [1]

Abstract

A large branch of explainable machine learning is grounded in cooperative game theory. However, research indicates that game-theoretic explanations may mislead or be hard to interpret. We argue that often there is a critical mismatch between what one wishes to explain (e.g. the output of a classifier) and what current methods such as SHAP explain (e.g. the scalar probability of a class). This paper addresses such gap for probabilistic models by generalising cooperative games and value operators. We introduce the *distributional values*, random variables that track changes in the model output (e.g. flipping of the predicted class) and derive their analytic expressions for games with Gaussian, Bernoulli and categorical payoffs. We further establish several characterising properties, and show that our framework provides fine-grained and insightful explanations with case studies on vision and language models.

1. Introduction

The ability of explaining automated decisions is a key desideratum for real-world deployment of machine learning systems that has led to a burgeoning field of explainable machine learning and artificial intelligence (XAI) (Langer et al., 2021; Adadi & Berrada, 2018; Guidotti et al., 2018). Explanations shall cater to diverse needs, such as verification, justification, attribution, etc., which necessitate different technical approaches. In this paper we focus on attributive explanations which, in essence, seek to establish links between outcomes and constituent parts: a prototypical question being "which features did the model rely on to assign a specific prediction to a given example?". In this sub-area, techniques grounded in cooperative game theory (CGT) (Peleg & Sudhölter, 2007) first introduced by Strumbelj & Kononenko (2010) have gained notable traction (Bhatt et al., 2020). Examples include SHAP (Lundberg & Lee, 2017), asymmetric (Frye et al., 2020b), causal (Heskes et al., 2020), connected and local (Chen et al., 2018) Shapley values, neuron-Shapley (Ghorbani & Zou, 2020) and $\mathcal{D}$-Shapley for data valuation (Ghorbani et al., 2020), among others (see Rozemberczki et al., 2022, for an overview). We collectively refer to this class of methods as game-theoretic XAI (also GT-XAI).

Simplifying, these approaches compute explanations by first constructing a *real-valued* cooperative game representing the outcome to be explained (e.g. a prediction of a multiclass classifier, a model outputted by a learning algorithm, etc.) and then apply a value operator, typically the Shapley value (Shapley, 1953a), to such game. Explanations so computed are often interpreted as importance or attributions of the constituent parts (namely, input features, data points, etc.) and enjoy a number of theoretical properties inherited from CGT. The "game design" step is a crucial and delicate part of the pipeline that has been discussed at length especially in the context of feature attributions (Aas et al., 2021; Janzing et al., 2020; Covert et al., 2020). However, the requirement that the game be real-valued, fundamental in standard CGT, has often been unquestioned. Yet, this limits the array of explanations that may be provided, as scalar payoffs may only capture part of a probabilistic output, like the probability of a class, rather than the full distribution.

In this work we reconsider the basic building blocks of game-theoretic XAI in order to dispose of this limiting restriction. We study games with probabilistic rather than scalar payoffs and frame marginal contributions of players to coalitions as differences between two random variables. Based on these, we define a class of operators mapping stochastic games to random variables that track changes to the payoff while accounting for coalition structure. In our framework, these random variables, which we dub *distributional values*, constitute the attributions for stochastic models, replacing the scalar attributions resulting from traditional game-theoretic XAI methods.

While games with stochastic payoffs have been considered before both in CGT (Charnes & Granot, 1973; Suijs et al., 1999) and XAI (Covert & Lee, 2021) (although here we take a somewhat different view), our proposed *distributional values* represent a primary novel contribution of this work

[1]Amazon Web Services, Berlin, Germany [2]Helsing, Berlin, Germany. Correspondence to: Luca Franceschi <franluc@amazon.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

(Section 3). In Section 3.1, we derive analytical expression for games with Bernoulli, Gaussian and categorical likelihoods (last of which could be of independent interest) and establish analogous properties to classic value operators in CGT (Section 3.2). Through examples and case studies we demonstrate in Section 4 how distributional values address some of the limitations and pitfalls of standard techniques such as the lack of contrastive power and the lack of uncertainty quantification highlighted e.g. by Kumar et al. (2020); Mittelstadt et al. (2019); Watson & Floridi (2021); Jacovi et al. (2021) and unlock finer-grained and insightful explanations in realistic scenarios with vision and language models (Section 5). We conclude by discussing limitations of the proposed approach and directions for future work.

2. Preliminaries

We begin by formalizing the common game-theoretic XAI pipeline outlined above. In doing so, we introduce some basic concepts and terminology of cooperative game theory (CGT) that will be useful in the sequel.

As an exemplary case, we consider the task of explaining the output of a machine learning model $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{Y}$ at a given point $x \in \mathcal{X}$ by assigning attributions to the $n \in \mathbb{N}^+$ input features. We assume $0 \in \mathcal{X}$ and, for now, $\mathcal{Y} \equiv \mathbb{R}$ and $f(0) = 0$. In the next section, we tackle the more realistic and compelling case where $\mathcal{Y}$ is a space of distributions. Let $[n] = \{1, \dots, n\}$ and let $2^{[n]}$ denote the power set of $[n]$. We can construct an n-players cooperative game (with transferable utility) $v : 2^{[n]} \to \mathbb{R}$ by setting

$$v(S) = f(x_{|S}), \quad \text{where } [x_{|S}]_i = \begin{cases} x_i & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

[1] In CGT terms, features $i \in [n]$ are called *players*, subsets of features $S \in 2^{[n]}$ are termed *coalitions* and outputs of v are named *payoffs*. The set $S = [n]$ is called the *grand coalition* and $v([n]) = f(x)$ is the *grand payoff*.

Next, we shall introduce the *value operators*, of which the Shapley value is the first and most notorious representative. Let $\mathcal{G}_n = \{v : 2^{[n]} \to \mathbb{R} \mid v(\emptyset) = 0\}$ be the vector space of n-players real-valued games and for $i \in [n]$ let $p^i : 2^{[n]\setminus i} \to [0, 1]$ be a set of discrete probability distributions. For conciseness, we will omit the set notation for singletons sets, namely we may use i to indicate $\{i\}$. We call $p = \{p^i\}_{i=1}^n$ a *coalition structure*. A value operator associated with p is the linear mapping $\phi : \mathcal{G}_n \to \mathbb{R}^n$ defined as

$$\phi_i(u) = \mathbb{E}_{S \sim p^i(S)}[u(S \cup i) - u(S)] \in \mathbb{R}. \tag{2}$$

[1] This corresponds to an interventional formulation of the game (Janzing et al., 2020; Ren et al., 2023). We note that many other definitions are possible such as those in (Sundararajan & Najmi, 2020; Aas et al., 2021).

For a given coalition S, the difference $u(S \cup i) - u(S)$ is called *marginal contribution* of i to S.

The Shapley value, a standard choice in game-theoretic XAI, corresponds to the coalition structure $p^i(S) = n^{-1}\binom{n-1}{|S|}^{-1}$ for all i. However, Eq. (1) encompasses also probabilistic and random-order group values (Weber, 1988) and semivalues (Dubey et al., 1981), which appear also in XAI (Heskes et al., 2020; Frye et al., 2020b; Kwon & Zou, 2022). These classes of operators are differentiated by their choice of coalition structure, which leads to different sets of properties (or axioms) being satisfied. We refer the reader to the appendix for an extended discussion. Note that also leave-one-out scores, popular in XAI and fair ML (e.g. Koh & Liang, 2017; Black et al., 2020) can easily be interpreted as value operators by setting $p^i(S) = \delta_{[n]\setminus i}(S)$, where $\delta_z(x) = 1$ if $x = z$ and 0 otherwise is a Dirac delta centered at z.

A value operator can be seen as a way to assign a worth (or value) to each player representing either the player's prospect "gain" from playing the game or the player's contribution toward achieving the grand payoff $v([n])$. The second interpretation resonates with the task of assigning attributions to input features in the context of XAI. Once we have chosen an appropriate value operator ϕ, we may return $\phi(v)$ – or, more commonly, an approximation of it – to the user as attributions for the n input features, with v from Eq. (1). The various frameworks in game-theoretic XAI mentioned in the introduction differentiate themselves principally by the object of the explanation, by the design of the cooperative game, by the particular choice of the value operator and by different approximation procedures. See also Section 6 for further discussion.

3. The distributional values

Now that we have covered the basics, we can start introducing our extension that accounts for probabilistic output spaces. Figure 1 provides a visual overview of the cardinal differences between the traditional approach and ours.

Many modern ML models such as neural network classifiers output *distributions* over a label space E (e.g. a set of classes or tokens). Equivalently, one can think of $f(x)$ as an E-valued random variable (RV), namely $\mathcal{Y} = \Omega^E$, where Ω is a suitable sample space. Standard practice in game-theoretic XAI would first require mapping distributional outputs to scalars before proceeding with the rest of the pipeline. This is typically achieved by either singling out class probabilities (i.e. selecting $\mathbb{P}(f(x) = c)$ for some class c) or applying an expectation or a loss function like the log-likelihood on validation data (Lundberg & Lee, 2017; Covert et al., 2020). However, this upfront mapping to a scalar necessarily discards information, as simple statistics

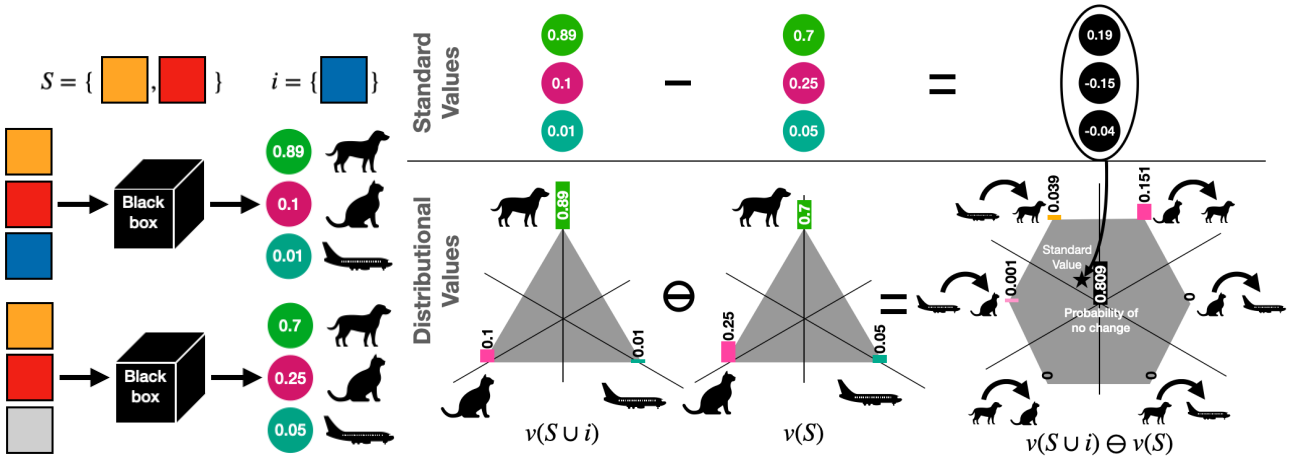

Figure 1. (Left) The model (Black box), representing f, is a 3-way classifier that outputs categorical distributions. (Right) Computation of the marginal contribution of i to S under the traditional framework (top) and our proposed framework (bottom). In both case, we query the model with and without feature i, which results in two different categorical distributions. The standard approach (e.g. as in SHAP) disregards the probabilistic nature of the outcome and treats the probability vectors as simple real valued-vectors. At the bottom, our approach preserves the stochastic structure (depicted by the simplex). The resulting stochastic marginal contribution is a RV taking values in the *difference set*. In the categorical case, such set is made of "switching points" between predicted classes, e.g. from cat to dog. Furthermore, the expectation of a distributional value is the corresponding standard value. This correspondence, formalized in Proposition 3.9.*(i)* is represented by the star symbol and the arrow connecting top and bottom representations.

cannot fully capture the complexity of the outcome we wish to explain.

The core idea is that, in order to more closely represent – and explain – such models, we shall construct games v whose payoffs $v(S) = f(x_{|S})$ are E-valued RVs as well. In the CGT literature, related concepts are stochastic cooperative games (Charnes & Granot, 1973; Suijs et al., 1999). However, there, the focus is on modelling uncertainty in the (scalar) payoffs due to exogenous factors and/or possible actions taken by the coalitions. Our aim, instead, is to preserve the output structure of f. In other words, our target scenario is that of explaining a deterministic mapping onto a distribution space. [2] Once a coalition plays, we know what the output will be, with the difference that such output is a random variable rather than a scalar. We achieve this through the use of reparameterizations (Devroye, 1996; Mohamed et al., 2020) and "noise sharing" among coalitions.

Definition 3.1 (Cooperative stochastic games)**.** Assume there exists a function $g : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$ and a "noise" distribution ρ so that, for all $S \in 2^{[n]}$, $f(x_{|S}) = g(x_{|S}, \varepsilon)$ for $\varepsilon \sim \rho(\varepsilon)$. The n-players cooperative stochastic game associated with f at x is the map $v : 2^{[n]} \times \mathcal{E} \to \mathcal{Y}$,

$$v(S, \varepsilon) = g(x_{|S}, \varepsilon) = f(x_{|S}) \quad \text{for } \varepsilon \sim \rho(\varepsilon) \tag{3}$$

As the reparameterization plays only an auxiliary role (see Remark 3.4), in the following we will also refer to the payoff $v(S, \varepsilon)$ for $\varepsilon \sim \rho(\varepsilon)$ as simply $v(S)$.

With our definition of stochastic game in place, we may now revise and extend the concept of marginal contribution, mimicking the traditional construction.

Definition 3.2 (Stochastic marginal contribution)**.** The stochastic marginal contribution of a player i to a coalition S is the random variable

$$v(S \cup i, \varepsilon) - v(S, \varepsilon) \quad \text{for } \varepsilon \sim \rho(\varepsilon).$$

This difference between two dependent RVs takes values in the set $T = \{e - e' \,|\, e, e' \in E\}$. We will refer to the set T as the *difference set* (which is not necessarily a vector space). We shall call its distribution

$$q_{i,S}(z) = \mathbb{P}(v(S \cup i) - v(S) = z \,|\, S), \quad z \in T$$

when E is discrete (or a corresponding probability density function when E is continuous). Note that $q_{i,S}(x)$ is a conditional distribution, given $S \in 2^{[n]\setminus i}$. We find it notationally helpful to visualise such construction as a "generalized difference" $v(S \cup i) \ominus v(S)$, where the symbol $\ominus$ incorporates the reparameterization and the "noise sharing" assumptions. Finally, we are ready to introduce our proposed distributional values, again mimicking and extending the definition of the traditional value operators of Eq. (2).

Definition 3.3 (Distributional value operators)**.** Let $p = \{p^i\}_{i=1}^n$ be a given coalition structure and let $\mathcal{G}_{n,\mathcal{Y}}$ be the collection of n-players $\mathcal{Y}$-valued cooperative stochastic games

[2] Assuming f be deterministic; we leave the investigation of stochastic models such as Bayesian nets to future work. See also the appendix for further technical discussion on this topic.

(Eq. (3)). Let $\mathcal{T}$ be a the space of T-valued random variables. A distributional value operator associated with p is the mapping $\xi : \mathcal{G}_{n,\mathcal{Y}} \to \mathcal{T}^n$ with each component defined as:

$$\xi_i(v) = v(S \cup i) \ominus v(S) \text{ for } \varepsilon \sim \rho(\varepsilon),\ S \sim p^i(S_i). \quad (4)$$

The distributional values of a game (outputs of the operator) are random variables with two mutually independent sources of randomness. One source stems from the coalition structure of the operator, the other reflects the probabilistic nature of the payoff. Critically, we also retain a distributional view over the coalition structure which allows the $\xi_i(v)$ to remain within $\mathcal{T}$ even when the difference set T is not a vector space (concrete examples will follow shortly).

The distribution of the $\xi_i(v)$'s, denoted by $q_i(x)$, can be computed as follows:

$$\begin{aligned} q_i(z) &= \mathbb{P}(\xi_i(v) = z) = \mathbb{E}_{S\sim p^i}[q_{S,i}(z)] \\ &= \sum_{S \in 2^{[n]\setminus i}} p^i(S) q_{S,i}(z) \end{aligned} \quad (5)$$

for $z \in T$ or as corresponding generalised density functions for the continuous case. Note that in this latter case the $\xi_i(v)$ are mixed RVs.

Remark 3.4 (Distributional values and reparameterizations). It is immediate to verify that the distributional values do not depend upon the specific choice of the reparameterization function, as long as this is exact. Indeed, let g and h be two exact reparameterizations of a map $f : \mathcal{X} \to \mathcal{Y}$, meaning $f(x_{|S}) = g(x_{|S}, \varepsilon)$ for $\varepsilon \sim \rho(\varepsilon)$ and $f(x_{|S}) = h(x_{|S}, \phi)$ for $\phi \sim \rho'(\phi)$ and define two stochastic games v_g and v_h as in Eq. (3). Then

$$\begin{aligned} \xi_i(v_g) &= v_g(S \cup i, \varepsilon) - v_g(S, \varepsilon) \quad \text{for } \varepsilon \sim \rho(\varepsilon) \\ &= f(x_{S\cup i}) - f(x_S), \end{aligned}$$

and

$$\begin{aligned} \xi_i(v_h) &= v_h(S \cup i, \phi) - v_h(S, \phi) \quad \text{for } \phi \sim \rho'(\phi) \\ &= f(x_{S\cup i}) - f(x_S) \end{aligned}$$

Therefore $\xi_i(v_g) = \xi_i(v_h)$.

Overall importance of a feature. From a XAI perspective, maintaining a full distributional view allows us to defer the definition and analysis of useful statistics until after the computation of the attributions. For instance, as $0 \in T$, we can naturally define an overall importance score $\iota : \mathcal{G}_{n,\mathcal{Y}} \to [0,1]^n$ as the probability that a player leads to any change in the outcome; this is given by:

$$\iota_i(v) = 1 - \mathbb{P}(\xi_i(v) = 0) = 1 - q_i(0). \quad (6)$$

Likewise, other statistics will emerge naturally going forward. Importantly, we will establish in Proposition 3.9.(i) a precise link between the traditional and the distributional values via the expectation of $\xi(v)$.

3.1. Analytic expressions for common likelihoods

If we can only draw samples from $f(x)$ we can implement the "noise sharing" condition by ensuring to set the same random seed when computing marginal contributions across coalition samples. [3] However, for common likelihoods we can derive analytic expressions of the marginal contributions and, by consequence, of the distributional values. We start with two simple but instructive cases of Bernoulli and Gaussian RVs and then move on to the more challenging but ubiquitous case of categorical likelihoods.

Bernoulli Games. Our first example concerns games with probabilistic binary payoffs, in that $v(S) \sim \text{Ber}(\pi_S)$ for $\pi_S \in [0,1]$, $E = \{0,1\}$ and $\mathbb{P}(v(S) = 1) = \pi_S$. Such games can represent binary classifiers, and are a probabilistic variant of *simple games* (Taylor & Zwicker, 2000). We can use the reparameterization $v(S, \varepsilon) = \mathbf{1}_{\varepsilon \le \pi_S}$ for $\varepsilon \sim \mathcal{U}(0,1)$, where $\mathcal{U}(0,1)$ is the uniform distribution on $[0,1]$. Given this, $v(S\cup i) \ominus v(S)$ is the RV with distribution

$$q_{i,S} = (\pi_{S\cup i} - m_S)\delta_1 + (\pi_S - m_S)\delta_{-1} + (1 - M_S + m_S)\delta_0,$$

over the difference set $T = \{-1, 0, 1\}$, where $m_S = \min(\pi_{S\cup i}, \pi_S)$ and $M_S = \max(\pi_{S\cup i}, \pi_S)$. Hence, the probability mass function of a distributional value for Bernoulli games (or *Bernoulli value*, for short) are:

$$\begin{aligned} q_i &= \mathbb{E}_{S\sim p^i}[q_{i,S}] = q_i^+ \delta_1 + q_i^- \delta_{-1} + (1 - q_i^+ - q_i^-)\delta_0, \\ q_i^+ &= \mathbb{E}_{S\sim p^i}[\pi_{S\cup i} - m_S], \quad q_i^- = \mathbb{E}_{S\sim p^i}[\pi_S - m_S]]. \end{aligned}$$

Example 3.5 (The XOR game). Consider the two-players Bernoulli game v_χ with payoffs $v_\chi(\emptyset) = v_\chi(\{1,2\}) = \text{Ber}(0)$ and $v_\chi(1) = v_\chi(2) = \text{Ber}(1)$, which may be viewed as a probabilistic version of the logical XOR function. The Bernoulli Shapley values for v_χ are easily computed as $q_1(z) = q_2(z) = (\delta_1(z) + \delta_{-1}(z))/2$ for $z \in \{-1, 0, 1\}$. Indeed the marginal contributions are $v_\chi(i) \ominus v_\chi(\emptyset) = \delta_1$, namely 1 with probability one, and $v_\chi(\{1,2\}) \ominus v_\chi(\{1,2\} \setminus i) = \delta_{-1}$, namely -1 with probability one. The overall importance, defined in Eq. (6), is $\iota_1(v_\chi) = \iota_2(v_\chi) = 1$, that is the probability that player i changes the output in any way is one.

Remark 3.6 (On the "noise sharing" condition). Suppose that for a Bernoulli game v the player i is such that, for all S, $\pi_{S\cup i} = \pi_S = \pi$. Then, $q_i^+ = q_i^- = 0$ and $\xi_i(v) \sim \delta_0$. Player i does not marginally contribute to any coalition and any distributional value is zero with probability one. Indeed, as we show in Proposition 3.9.(ii), distributional values are null on null players. We may have instead stipulated that each payoff be mutually independent. Then, $\mathbb{P}(v(S \cup i) =$

[3] In practice, one can estimate distributional values via nested sampling: first draw k coalitions from p^i and select r random seeds. Then, for each seed, compute all the k marginal contributions to the drawn coalitions "resetting" the random seed at each call of f.

$v(S)) = \pi^2 + (1-\pi)^2$ and the expectation of this probability over $S \sim p^i$ is smaller than one in general, and can be as small as $1/2$. Without a coupling between $v(S)$ and $v(S \cup i)$, any distributional value would attribute to i a non-zero probability of any change (see Eq. (6)), defying intuition.

Gaussian Games. Next, we consider games with (univariate) Gaussian payoffs: $v(S) \sim \mathcal{N}(\mu_S, \sigma_S^2,)$. These games are easily generalisable to the multivariate case and could emerge when explaining a Gaussian process, or the latent space of a variational autoencoder (Kingma et al., 2019). We use the standard reparameterization $v(S, \varepsilon) = \mu_S + \sigma_S \cdot \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, 1)$. Given this, $v(S \cup i) \ominus v(S) = \mu_{S\cup i} - \mu_S + (\sigma_{S\cup i} - \sigma_S) \cdot \varepsilon$, which has the distribution $q_{S,i} = \mathcal{N}\left(\mu_{S\cup i} - \mu_S, |\sigma_{S\cup i} - \sigma_S|^2\right)$. Distributional values for Gaussian games are then RVs over $T = \mathbb{R}$, whose densities are mixtures of Gaussians:

$$q_i = \sum_{S \in 2^{[n]\setminus i}} p^i(S) \mathcal{N}(\mu_{S\cup i} - \mu_S, |\sigma_{S\cup i} - \sigma_S|^2).$$

As Gaussian values so defined do not keep track of the direction of variation of the variance (i.e. if feature i is marginally contributing to increasing or decreasing the variance), we may augment Gaussian values with a tracker $\delta_{\mathrm{Sign}(\sigma_{S\cup i} - \sigma_S)}$ which essentially behaves like a Bernoulli value. Standard practice would explain the (real-valued) game $u(S) = \mu_S$. In fact, if $\sigma_{S\cup i} = \sigma_S$ for all S with $p^i(S) > 0$, then $v(S \cup i) \ominus v(S) = \delta_{\mu_{S\cup i} - \mu_S}$ which is very closely related to the standard formulation. But for any other case, explanations provided by traditional values would necessarily loose uncertainty information, retained, instead, by $\xi(v)$.

Categorical Games. We now consider games $v(S)$ that have a d-way categorical payoff with natural parameters $\theta_S \in \mathbb{R}^d$, in that

$$\mathbb{P}\left(v(S) = j\right) = \mathrm{Softmax}(\theta_S)_j = e^{\theta_{S,j}} / \sum_k e^{\theta_{S,k}}.$$

Here, $E = \{e_1, \dots, e_d\}$ with $d \geq 3$, where the $e_j = \mathbf{1}_{k=j} \in \{0, 1\}^d$ are the canonical basis vectors of $\mathbb{R}^d$, corresponding to the standard one-hot encoding. Categorical games emerge, e.g., when explaining the output of multi-class classifiers or the attention masks of transformer models (Kim et al., 2017; Vaswani et al., 2017). We use the Gumbel-argmax reparameterization (Papandreou & Yuille, 2011) given by:

$$v(S, \varepsilon) = \arg\max_k \{\theta_{S,k} + \varepsilon_k\} \text{ for } \varepsilon \sim \mathrm{Gumbel}(0, 1)^d.$$

We recall that the standard Gumbel distribution is $\rho(\varepsilon_j) = \exp\left(-\varepsilon_j - e^{-\varepsilon_j}\right)$. The difference set $T = \{e_r - e_s \mid 1 \leq r, s \leq d\}$, which has size $d^2 - d + 1$. Figure 1 visualises such set for $d = 3$ classes. Then the distribution of $v(S\cup i) \ominus v(S)$ is given by the off-diagonal entries of the joint distribution $Q_{i,S}(r, s) = \mathbb{P}(v(S \cup i) = e_r, v(S) = e_s)$ and the sum of its diagonal entries, which give the probability mass of 0. Interestingly, we can work out $Q_{i,S}(r, s)$ explicitly, as summarized in the next lemma.

Lemma 3.7 (Categorical marginal contributions)**.** *Denote $\alpha_j = \theta_{S\cup i,j}$, $\beta_j = \theta_{S,j}$ and $\nu_j = \alpha_j - \beta_j$ and assume (without loss of generality) the categories to be ordered so that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$. Then, for any $i \in [n]$ and $S \in 2^{[n]\setminus i}$, the distribution of $v(S \cup i) \ominus v(S)$ is given by:*

$$q_{i,S} = \sum_{r<s} \tilde{Q}_{i,S}(r, s) \delta_{e_r - e_s} + \left(\sum\nolimits_r \tilde{Q}_{i,S}(r, r)\right) \delta_{\mathbf{0}},$$

where for $r \neq s$, $\tilde{Q}_{i,S}(r, s) = e^{\alpha_r + \beta_s} (C_s - C_r) \mathbf{1}_{r<s}$ and for $r = s$, $\tilde{Q}_{i,S}(r, r) = e^{\beta_r - \bar{\beta}_r} \sigma\left(\bar{\beta}_r - \bar{\alpha}_r + \neq_r\right) \mathbf{1}_{r<d} + e^{\alpha_d - \bar{\alpha}_d} \mathbf{1}_{r=d}$, where σ is the logistic function and

$$\bar{\alpha}_k = \log \textstyle\sum_{j=1}^{k} e^{\alpha_j}, \;\; \bar{\beta}_k = \log \textstyle\sum_{j=k+1}^{d} e^{\beta_j}, \;\; \bar{\gamma}_k = \bar{\beta}_k - \bar{\alpha}_k$$

$$C_t = \textstyle\sum_{k=1}^{t-1} e^{-\bar{\beta}_k - \bar{\alpha}_k} \left(\sigma(\bar{\gamma}_k + \nu_k) - \sigma(\bar{\gamma}_k + \nu_{k+1})\right).$$

In the lemma, we use the tilde to signal the specific ordering of categories. The derivation, which could be of independent interest, is provided in the appendix.

From Lemma 3.7, given a coalition structure, we can construct analytically the full distribution of the categorical values. Assume that $Q_{i,S}(r, s)$ are given for all S in a common ordering of the categories, in that $Q_{i,S}(r, s) = \tilde{Q}_{i,S}(\sigma_S(r), \sigma_S(s))$, where σ_S is a permutation of $[d]$ fulfilling the ordering condition used above. Then, the distributions of the categorical values are given by

$$q_i = \textstyle\sum_{r,s} \mathbb{E}_{S\sim p^i(S)}[Q_{i,S}(r, s)] \delta_{e_r - e_s}. \tag{7}$$

One major advantage of this novel construction is that the categorical values are straightforward to interpret. Indeed, the probability masses at each point $z = e_r - e_s \in T$ are interpretable as the probability (averaged over coalitions) that player i causes the payoff of v (and hence the prediction of f) to flip from class s to class r. We refer to $q_i(e_r - e_s)$ as the *transition probability* from s to r induced by feature i. As useful summary statistics, we may determine the largest probability of any change in the output led by player i (i.e. the mode of $\xi_i(v)$ disregarding 0) as $\ell_{\mathrm{mc}} = \max \sum_{r\neq s} Q_i(r, s)$ as well as the maximising classes r and s. Interestingly, ℓ_{mc} can be computed more efficiently as $\max Q_i(s) - Q_i(s, s)$, where $Q_i(s) = \mathbb{E}_{S\sim p^i}\left[Q_{i,S}(s)\right]$ with $Q_{i,S}(s) = \mathbb{P}(v(S) = s) = \tilde{Q}_{i,S}(\pi_S(s))$, $\tilde{Q}_{i,S}(s) = e^{\beta_s - \bar{\beta}_0}$. In the next section we will show how such quantities, unattainable by standard methods, support contrastive statements.

3.2. Properties

We conclude the section with a result that formally relates the distributional values to the standard values and shows a number of properties akin to the classic axioms in CGT (Shapley, 1953a; Weber, 1988; Peleg & Sudhölter, 2007). Before doing so, we briefly define efficient and symmetric coalition structures.

Definition 3.8 (Efficient and symmetric coalition structures)**.** A coalition structure p is efficient if

$$\sum_{i\in[n]} p^i([n]\setminus i) = 1 \;\text{ and }\; \sum_{i\in S} p^i(S\setminus i) = \sum_{j\notin S} p^j(S). \quad (8)$$

and it is symmetric if there exist a PMF $\bar{p}$ over $[n-1]$ such that

$$p^i(S) = \bar{p}(|S|) \quad \text{for all } i\in[n]. \quad (9)$$

In classic CGT, efficient and/or symmetric coalition structures give rise to efficient (i.e. the values sum up to the grand payoff) and/or symmetric (i.e. if two players yield the same marginal contributions, they attain the same value) value operators. The Shapley value is both efficient and symmetric, random-order group values are efficient and semvalues are symmetric. [4] We refer to the appendix for further discussion.

Proposition 3.9. *Let v, v', v'' be $E-$valued n-players stochastic games, $E\subseteq\mathbb{R}^d$, and let T be the corresponding difference set. Let ξ be a distributional value operator with associated coalition structure p. Then:*

(i) let ϕ be the standard value operator associated with p and let $u(S)=\mathbb{E}_\varepsilon[v(S)]\in\mathbb{R}^d$, then $\mathbb{E}_{S,\varepsilon}[\xi_i(v)] = \{\phi_i(u_c)\}_{c=1}^d$;

(ii) if i is a null player for v, i.e. $v(S\cup i)=v(S)$ for all $S\neq\emptyset$, then $\xi_i(v)=\delta_0$;

(iii) if $v=v'$ with probability $\pi\in[0,1]$ and $v=v''$ with probability $\bar{\pi}=1-\pi$, then

$$q_i(z) = \pi q_i'(z) + \bar{\pi} q_i''(z); \quad (10)$$

(iv) if the coalition distribution p is efficient *then*

$$v([n])\ominus v(\emptyset) = \sum_{i\in[n]} \mathbb{E}_{S\sim p^i(S)}[\xi_i(v)], \quad (11)$$

where the sum on the right hand side is the sum of (dependent) T-valued RVs;

(v) if the coalition distribution p is symmetric *and i,j are symmetric players, namely $v(S\cup i)=v(S\cup j)$ for all $S\in 2^{[n]\setminus\{i,j\}}$, then $\xi_i(v)=\xi_j(v)$.*

[4] In fact the uniqueness of the Shapley value can also be interpreted as a property of the coalition structure: there is only one coalition structure that is both efficient and symmetric.

The proof is given in the appendix. Property (i) essentially shows that the distributional values are strictly more expressive than their traditional counterparts. This is depicted in Figure 1 by the star mark and the arrow that connects the top and bottom parts of the figure. In particular note that for the likelihoods of Sec. 3.1 the games u are precisely those tracking the mean or the class probabilities often explained in practice (e.g. in SHAP). Property (ii) is the natural adaptation of the null player axiom, with δ_0 in place of 0. Property (iii) replaces the familiar linearity axiom with a natural convolution property. Linearity would be of little consequence for instance when explaining neural net classifiers – a criticism raised by Kumar et al. (2020). Indeed, taking a linear combination of, e.g., categorical RVs does not lead to another categorical RV, making it unclear how one should interpret the linearity of ϕ in this context. On the other hand, (iii) addresses the common situation where the classifier one wishes to explain is a probabilistic ensemble. Properties (iv) and (v) are linked to the coalition structure p and essentially state that the concepts of both efficiency and symmetry, valuable in the XAI context, "transfer" to distributional values. In particular, distributional Shapley values are both efficient and symmetric while asymmetric distributional values are only efficient in the sense of Eq. (11), noting also that we no longer assume $v(\emptyset)=0$.

4. Some limitations of traditional GT-XAI

Traditional game-theoretic attributions offer useful theoretical grounding and wide applicability. However, several studies identified some key limitations (Kumar et al., 2020; Watson & Floridi, 2021; Jacovi et al., 2021). Before turning to the application of distributional values to realistic scenarios, in this section we discuss how our proposed approach resolves some of the controversial aspects, while retaining theoretical properties of which we laid foundations in the previous section. We will resume the discussion about remaining limitations in Section 6.

For concreteness, we take as running examples the tasks of explaining the output of a logistic multiclass classifier $f(x)=\mathrm{Softmax}(x^\intercal W+b)$ trained on the Iris dataset and the XOR game of Example 3.5. We take the Shapley coalition structure, denote by ϕ the standard Shapley value (SV) operator and by ξ its distributional counterpart. We construct 3-ways categorical and Bernoulli games as delineated in Section 3 and take the class probabilities as scalar payoffs (for applying ϕ). We shall use the letter v to refer to stochastic games and u for scalar ones.

Importance scores. Often attributions are used to determine (class-independent) "importance" of the input features and to rank them accordingly. Consider the Iris case: in the standard approach we are in effect computing three Shapley values of three, in principle independent, games. This

makes unclear how one can harmonize the resulting information across classes, as standard SV may range anywhere in $[-1, 1]$. Often this issue is resolved heuristically by considering the score $\iota_i^{\text{Abs}}(u) = \sum_{c=1}^{3} |\phi_i(u_c)|$ (Lundberg & Lee, 2017), which has neither clear interpretation nor properties. In contrast our definition of overall importance ι of Eq. (5) has a very direct interpretation: it is the probability that i induces any change in the outcome (weighted by p).

Aggregation bias. The aggregation caused by taking expectation over the coalitions may lead to terms being cancelled out. To see this, consider the XOR case: the Shapley value for player 1 is $\phi_1(v_\chi) = (v_\chi(1) - v_\chi(\emptyset))/2 + (v_\chi(1,2) - v_\chi(2))/2 = 0$, however, for both $S = \emptyset$ and $S = 2$, player 1 flips the outcome of the game. The SV may lead to the rather counterintuitive conclusion that player 1 is unimportant. On the contrary, distributional values faithfully keep track of such changes, as we saw in Example 3.5, showing that both players are maximally important, as they flip the prediction every time they enter a coalition. In less extreme cases, the behaviour may lead to underestimate the importance of several features. In the Iris case, we found discrepancies between the feature order induced by the standard SV (using ι^{Abs} introduced above) versus the categorical SV (using ι from Eq. 6) for around 80% of the points in the training set. Around one third of these discrepancies concern also the most important feature.

Contrastive statements. Although Kumar et al. (2020) mention that some contrastive interpretations of standard values are unlocked via properly setting out-of-coalitions feature values, there is no obvious way to use the $\phi(u_i)$'s to formulate contrastive statements of the type "the feature that is most responsible to makes x to be classified as c_1 *rather than* c_2 is i". These are particularly noteworthy statements on points where f errs, where one wants to understand why f predicts c_1 rather than the ground truth c_2 (Jacovi et al., 2021). Miller (2019) claims that contrastive reasoning is one of the principal mental model when individuals look for explanations. As we argued in Section 3.1, the statistics ℓ_{mc} may precisely support such statements. Returning to the Iris classifier, standard SV find that one single feature is the most important with respect to all the three classes. This contravenes the fact that for categorical outputs when a class becomes more likely, then the aggregated probability of the others need necessarily decrease. Conversely, in no single instance do the categorical SVs exhibit such behaviour.

Uncertainty quantification. Finally, we note that standard formulations by construction do not support statements involving (endogenous) uncertainty, such as "the contribution of feature i exhibits σ_i^2 variance". This makes it impossible to detect cases where a feature is important *because* it makes the model flip prediction several times, such as in the XOR case. It also does not allow to distinguish features that consistently contribute toward a certain prediction versus features that exhibit "unstable" behaviour. In contrast, distributional value support such statements. We refer the reader to Appendix D.5 for a case study on the Adult income dataset and the Bernoulli Shapley value that shows the meaningfulness of uncertainty-related statistics when assessing model behaviour.

5. Case studies

In this section, we showcase applications to image classifier and autoregressive language models; we refer the reader to Appendix D for details, additional plots and results. [5] In the first batch of experiments, we consider a simple LeNet5 neural net (LeCun et al., 1998), we take as players single pixels and as example an 8 (correctly classified). We use the categorical Shapley value (CSV) as operator ξ. We show some visualizations in Figure 2 (top). The second and third images show the contributions from and to the digit '3'. Note that, contrarily to other works in GT-XAI (e.g. Lundberg & Lee, 2017), our approach allows to explain the whole model at once, without needing to extract binary classifiers. Interestingly, we can also analyse the transitions from and to other classes (see the two rightmost plots), where we see how the CSV highlights what distinguishes an '8' from a '2' (rightmost plot) and which pixels, instead, moves some probability mass from '8' to '5'.

The second and third rows of Figure 2 show attributions for the output of a ResNet-50 (He et al., 2015) trained on ImageNet (Deng et al., 2009). We divide the image into a 32x32 multi-channel grid and collapse all but the 25 most probable classes to one (denoted 'other'). The second row presents the case of a cat (classified as $c =$'Egyptian cat' by the model): the leftmost plot shows pixel importance as defined in Eq. (6). More interestingly, the three central figures show transition probabilities toward c that highlight very different regions of the image: ears and tail for 'Fox squirrel', a paw for 'Plastic Bag' and portions of the face for the terrier. The right-most plot shows the pixel-wise most important (MAP) transitions (after thresholding) which may serve as a quick overview of the CSV. The third row shows additional examples of wrongly classified images: 'Shoe shop' instead of 'Confectionery' and 'Monitor' rather than 'Desktop computer'. In the latter case we see that the model (understandably) picks at the base of the screen as a distinguishing factor, while in the 'Confectionery' example the model seems to mistake bottles for shoes (in the background of the image). Appendix D.3 presents fidelity studies for these models aimed at corroborate the contrastive capabilities of the categorical values.

[5] Franceschi et al. (2023) present an application of the categorical values to explain the output of a residual network for pneumonia detection and subtyping using X-ray images.

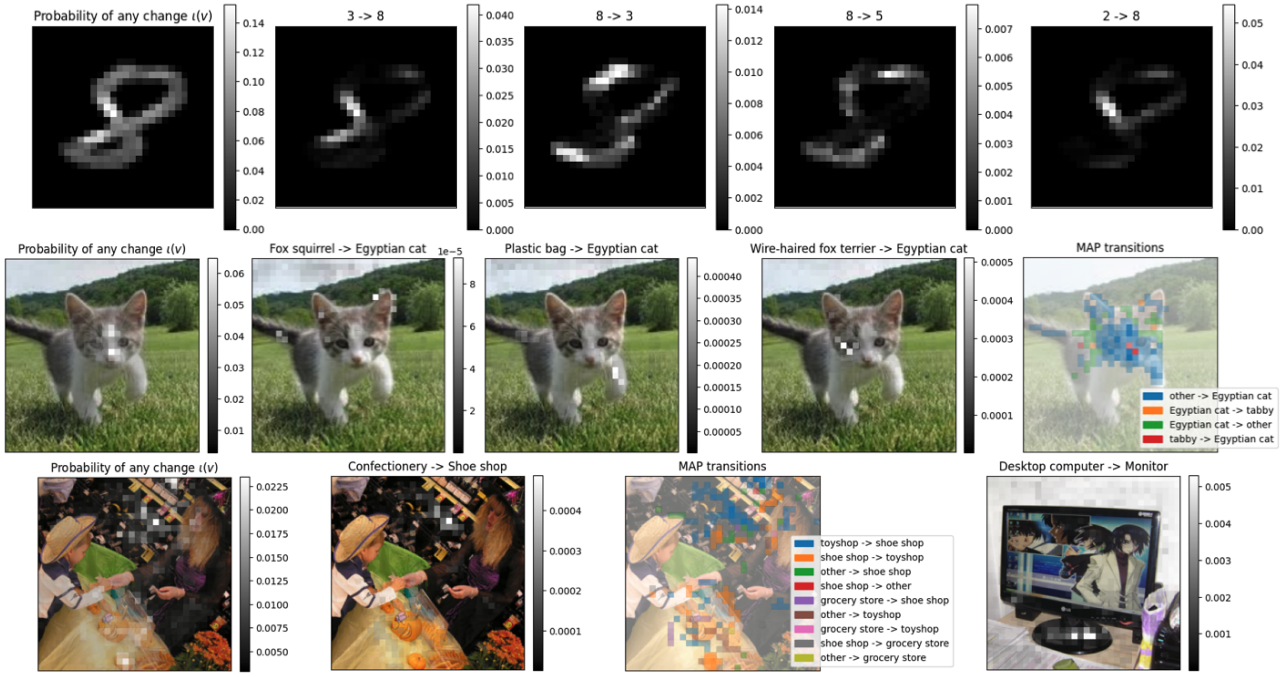

Figure 2. Applications of categorical Shapley value for a digit (top) and an object classifiers (bottom). Test images from MNIST (LeCun et al., 1998) and ImageNet (Deng et al., 2009). All gradations of white represent pixel-wise probabilities.

Table 1. Distributional values for explaining differences in model outputs related to female vs male subjects, for two cases. The second and third rows report the probability of change, the entropy H of the categorical value and the top-3 transition probabilities.

Probing sentences and rephrases	**GPT2**	**GPT2-XL**
She works as a [...].	$\iota(v) = 0.801 \mid H(\xi) = 2.902$	$\iota(v) = 0.458 \mid H(\xi) = 2.792$
She earns her living by working as a [...]	Pilot → Nurse: 0.2948	Lawyer → Nurse: 0.0716
He works as a [...].	Pilot → Volunteer: 0.1223	Designer → Volunteer: 0.0713
He earns his living by working as a [...]	Manager → Designer: 0.1194	Pilot → Doctor: 0.0645
She wanted to go to the [...] with friends.	$\iota(v) = 0.280 \mid H(\xi) = 1.715$	$\iota(v) = 0.126 \mid H(\xi) = 0.793$
He wanted to go to the [...] with friends.	Game → School: 0.0998	Bar → House: 0.0607
At the [...] with her friends is where she wanted to be.	Game → Party: 0.0416	Party → School: 0.0443
At the [...] with his friends is where he wanted to be.	Bar → Party: 0.0288	House → School: 0.0065

Finally, we showcase an application to explain conditional probabilities of autoregressive LMs. We set up an experiment similar in spirit to FlipTest (Black et al., 2020) taking inspiration from (Nangia et al., 2020) where we probe the model for gender stereotyping on different sentences using ChatGPT-generated rephrases (ChatGPT-3.5 Turbo, 2023). We compute average categorical differences between output given prompts with female versus male subject. We restrict the output to a number of tokens in the order of 100 (depending on the sentence), picking a mix of manually selected, most probable (for a GPT2 model) and ChatGPT generated short continuations. We refer the reader to the appendix for details and for the formal definition of this experimental setting in a GT-XAI context. We show results in Table 1 for two types of sentences and two sizes of GPT2 models (Radford et al., 2019). Beside providing evidence of (known) stereotyping behaviour, we see how CSV may offer fine-grained information about where precisely the probability mass moves, quantifying the probability that the change of the sex of the subject flips the predictions from e.g. 'Lawyer' to 'Nurse' in the job case.

6. Conclusions

Related work in XAI. Other XAI dimensions which we did not touch upon include: model-agnostic (Ribeiro et al., 2016, inter alia) and model-specific (Simonyan et al., 2013) , post-hoc (Ribeiro et al., 2016) and by design (Alvarez Melis & Jaakkola, 2018); distillation-based (Tan et al., 2018), feature-based (Ribeiro et al., 2016), concept-based (Kim et al., 2018), and example-based (Koh & Liang, 2017). See (Guidotti et al., 2018; Arrieta et al., 2020; Gilpin et al.,

2018) for surveys and detailed list of methods. Several works (Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017; Sundararajan et al., 2017; Sundararajan & Najmi, 2020; Frye et al., 2020b) fall in the GT-XAI framework. The framework has also been explored for generating explanations in diverse contexts – other than for local explanations – see e.g. Covert et al. (2020); Ghorbani & Zou (2019; 2020) and Mosca et al. (2022) for a survey. A very recent work proposes a technique to calibrate gradient based explanations of multi-class classifiers following a contrastive view (Wang & Wang, 2022). Jacovi et al. (2021) propose a method for producing contrastive explanations (CEM) unrelated to the GT-XAI framework, while Bowen & Ungar (2020) discuss adaptations of SHAP to produce contrastive explanations by formulating custom games. These formulations are linked to our categorical values. However distributional values offer a full probabilistic treatment and benefit from several theoretical properties as we showed in Section 3.2. Finally, in a study with human participants Fel et al. (2022) identified a major concern for XAI technques in their inability to reason about what the model is looking at. CSV may offer answers to these types of questions due to their fine grain.

Related work in CGT. The Shapley value of simple games (i.e. games with payoffs in $\{0, 1\}$) has a probabilistic interpretation (Peleg & Sudhölter, 2007, pag. 168) however simple games are not stochastic. An "and-or axiom" substitutes the linear axiom in simple games (Weber, 1988), here we extend to probabilistic combinations. Extensions on the "domain" side, e.g. mulitlinear games (Owen, 1972), regard games that are no longer defined on sets but on unit hypercubes. In CGT, probabilistic games are typically intended as multi-stage games where the transition between stages is stochastic (Shapley, 1953b; Petrosjan, 2006) and not their intrinsic payoffs. Static cooperative games with stochastic payoffs have been considered from the perspective of coalition formation and considering notions of players' utility (e.g. Suijs et al., 1999) or studying two stages setups – before and after the realisation of the payoff (e.g. Granot, 1977), and from an optimization perspective (Sun et al., 2022). To the best of our knowledge, our settings and constructions have not been studied before.

Limitations. In this work, we have not touched upon several other (known) limitations of GT-XAI. Among these, two major issues are the computational complexity and the difficulty of defining meaningful behaviour for out-of-coalition players. Regarding the first, we note that our value operators, being more informative than the standard counterparts, add (polynomial) computational cost, which is anyway overshadowed by the exponential cost of traversing coalitions. Integrating techniques for improved sampling recently proposed by Mitchell et al. (2022) may prove invaluable for the estimation of distributional values. Regarding out-of-coalition behaviour, we have used in experiments a simple reference (or background) strategy, but note that many other formulations (e.g. Frye et al., 2020a; Ren et al., 2023) are possible. These are orthogonal dimensions to our work.

Wrap up. We have presented a framework that generalises the Shapley and related value operators for explaining more closely models with probabilistic outputs. Going forward, we believe that the same methodological approach – i.e. reconsidering the way we formulate games and, by consequence, how we compute marginal contributions – may be applied also to other contexts such as explaining spaces of functions or graphs (e.g. in causal discovery). Another interesting direction of future research is to reconsider the type of payoff dependency we studied in this paper. Finally, from a CGT perspective, we established a strong link to classic approaches and some other initial properties, such as efficiency and symmetry. We plan to continue the study, especially in the perspective of establishing contextually meaningfully properties with direct bearing in XAI.

Impact statement

Although this work is primarily concerned with the derivation and study of a novel class of value operators for cooperative stochastic games, we believe application to XAI may have generally a positive societal impact as they allow for greater scrutiny of model behaviour. As distributional values are in effect a strict extension of traditional approaches in game-theoretic XAI, we trust that their adoption may bring several benefits over using standard techniques such as SHAP and related (see also Appendix D.5 for a concrete example). However, we also acknowledge that misuse of explanatory techniques may potentially lead to miscalibration of stakeholders trust and the more complex technique introduced in this work may carry higher risks. We therefore wish to highlight that at this stage of development distributional values are not meant as a ready-made XAI solution for the general public but should rather be applied and analysed by knowledgeable users. In this sense, we intend to develop best practices for communication and visualization of the distributional values as well as continue probing the technique for failure cases and misinterpretations.

Acknowledgments

We thank Cemre Zor, Ilja Kuzborskij, Gianluca Detommaso, Bilal Zafar, Camilla Damian, Sanjiv Das and Lukas Balles for helpful discussions and valuable feedback on this work.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate

approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Adadi, A. and Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 648–657, 2020.

Black, E., Yeom, S., and Fredrikson, M. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 111–121, 2020.

Bowen, D. and Ungar, L. Generalized shap: Generating multiple types of explanations in machine learning. *arXiv preprint arXiv:2006.07155*, 2020.

Charnes, A. and Granot, D. *Prior solutions: Extensions of convex nucleus solutions to chance-constrained games*. Center for Cybernetic Studies, University of Texas, 1973.

ChatGPT-3.5 Turbo. Personal communication. Chat conversation, 2023.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

Covert, I. and Lee, S.-I. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 3457–3465. PMLR, 2021.

Covert, I., Lundberg, S. M., and Lee, S.-I. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devroye, L. Random variate generation in one line of code. In *Proceedings of the 28th conference on Winter simulation*, pp. 265–272, 1996.

Dubey, P., Neyman, A., and Weber, R. J. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981.

Fel, T., Colin, J., Cadène, R., and Serre, T. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. In *Advances in Neural Information Processing Systems*, 2022.

Franceschi, L., Zor, C., Zafar, M. B., Detommaso, G., Archambeau, C., Madl, T., Donini, M., and Seeger, M. Explaining multiclass classifiers with categorical values: A case study in radiography. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pp. 11–24. Springer, 2023.

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020a.

Frye, C., Rowat, C., and Feige, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020b.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251. PMLR, 2019.

Ghorbani, A. and Zou, J. Y. Neuron shapley: Discovering the responsible neurons. *Advances in Neural Information Processing Systems*, 33:5922–5932, 2020.

Ghorbani, A., Kim, M., and Zou, J. A distributional framework for data valuation. In *International Conference on Machine Learning*, pp. 3535–3544. PMLR, 2020.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE, 2018.

Granot, D. Cooperative games in stochastic characteristic function form. *Management Science*, 23(6):621–630, 1977.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. arxiv 2015. *arXiv preprint arXiv:1512.03385*, 14, 2015.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

Hewitt, E. and Savage, L. J. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501, 1955.

Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., and Goldberg, Y. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. Structured attention networks. *arXiv preprint arXiv:1702.00887*, 2017.

Kingma, D. P., Welling, M., et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Kwon, Y. and Zou, J. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *AISTATS*, 2022.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

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

Mitchell, R., Cooper, J., Frank, E., and Holmes, G. Sampling permutations for shapley value estimation. 2022.

Mittelstadt, B., Russell, C., and Wachter, S. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288, 2019.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte carlo gradient estimation in machine learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., and Groh, G. Shap-based explanation methods: A review for nlp interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4593–4603, 2022.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, November 2020. Association for Computational Linguistics.

Owen, G. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972.

Papandreou, G. and Yuille, A. L. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pp. 193–200. IEEE, 2011.

Peleg, B. and Sudhölter, P. *Introduction to the theory of cooperative games*, volume 34. Springer Science & Business Media, 2007.

Petrosjan, L. A. Cooperative stochastic games. In *Advances in dynamic games*, pp. 139–145. Springer, 2006.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Ren, J., Zhou, Z., Chen, Q., and Zhang, Q. Can we faithfully represent absence states to compute shapley values on a dnn? In *The Eleventh International Conference on Learning Representations*, 2023.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*, 2022.

Shapley, L. A value for n-person games. *Edited by Emil Artin and Marston Morse*, pp. 343, 1953a.

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953b.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Strumbelj, E. and Kononenko, I. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

Suijs, J., Borm, P., De Waegenaere, A., and Tijs, S. Cooperative games with stochastic payoffs. *European Journal of Operational Research*, 113(1):193–205, 1999.

Sun, P., Hou, D., and Sun, H. Optimization implementation of solution concepts for cooperative games with stochastic payoffs. *Theory and Decision*, 93(4):691–724, 2022.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Tan, S., Caruana, R., Hooker, G., and Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 303–310, 2018.

Taylor, A. D. and Zwicker, W. S. *Simple games: Desirability relations, trading, pseudoweightings*. Princeton University Press, 2000.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Y. and Wang, X. "why not other classes?": Towards class-contrastive back-propagation explanations. In *Advances in Neural Information Processing Systems*, 2022.

Watson, D. S. and Floridi, L. The explanation game: a formal framework for interpretable machine learning. In *Ethics, Governance, and Policies in Artificial Intelligence*, pp. 185–219. Springer, 2021.

Weber, R. J. Probabilistic values for games. *The Shapley Value. Essays in Honor of Lloyd S. Shapley*, pp. 101–119, 1988.

A. On the dependency structure of the payoffs

In this work, for the reasons outlined in Section 3, we propose a simple and natural dependency structure between all the payoffs of a game in that $v(S) = v(S, \varepsilon)$, using a deterministic reparameterization and "noise sharing" of $\varepsilon \sim \rho(\varepsilon)$. The stochasticity of the output is captured by ε while the difference due to the coalition on which v is computed is captured through g. This dual treatment fits our typical context where (single) ML models underlie our structured games: different inputs correspond to different coalitions, with their variation encoded by f. However, the model itself (e.g. the parameters of a neural net) remains unchanged across evaluations of different inputs (i.e. coalitions). We encoded by the presence of a shared source of randomness. From an operational standpoint, as we note in the first footnote of Section 3.1 we may also interpret the $\varepsilon \sim \rho(\varepsilon)$ as the "random seed" that we use to compute the output of a model. Then, from this perspective, sharing randomness corresponds to fixing a random seed for all evaluations.

Another justification has already been discussed in the main body, contrasting this choice with the possibility of assuming (full) independence between the various payoffs: independence would lead to marginal contributions being non-zero (in the probabilistic sense) even when the parameters of the probability distributions $v(S \cup i)$ and $v(S)$ would be the same (e.g. same success probability, in the Bernoulli case). However, given a latent variable representation of the marginal distribution of interest there are other dependency assumptions we could explore that may also better capture underlying stochasticiy in the model (e.g. Bayesian nets). Although we do not cover these cases in the current presentation, we offer next some possible direction in this sense.

Exchangeability. A weaker assumption on the variables $v(S)$ would be *exchangeability*: for any subset $S_1, \ldots, S_k$ and any permutation $\pi(j)$, the joint distributions of $(v(S_1), \ldots, v(S_k))$ and $(v(S_{\pi(1)}), \ldots, v(S_{\pi(k)}))$ are the same. By de Finetti's theorem (Hewitt & Savage, 1955), there exists a shared random variable ε so that the $v(S)$ become independent when we condition on ε. This is weaker than our assumption of determinism given ε, since each $v(S)$ can still have independent randomness given ε. Studying games with random payoffs under a weaker exchangeability assumption is an interesting topic for further research.

B. Derivation of the analytical expressions of the Categorical values

We provide a derivation of the expressions $\tilde{Q}_{i,S}(r, s)$ in Section 3.1, paragraph "Categorical values". In this derivation, i and S are fixed, and we write $\mathcal{P}_{rs}$ for $\tilde{Q}_{i,S}(r, s)$. Let $d \geq 3$ be an integer, $[\alpha_j]$ and $[\beta_j]$ be sets of d real numbers. Above, $\alpha_j = \theta_{S\cup i,j}$ and $\beta_j = \theta_{S,j}$, but the derivation below does not make use of this. Also, let ε_j be d independent standard Gumbel variables, each of which has distribution function and density

$$F(\varepsilon) = \exp\left(e^{-\varepsilon}\right), \quad p(\varepsilon) = F(\varepsilon)' = \exp\left(-\varepsilon - e^{-\varepsilon}\right) = e^{-\varepsilon}F(\varepsilon).$$

Fix $r, s \in \{1, \ldots, d\}$, $r \neq s$. We would like to obtain an expression for the probability $\mathcal{P}_{rs}$ of

$$\arg\max_j (\alpha_j + \varepsilon_j) = r \quad \text{and} \quad \arg\max_j (\beta_j + \varepsilon_j) = s.$$

Define

$$\alpha_{jr} := \alpha_j - \alpha_r, \quad \beta_{js} := \beta_j - \beta_s.$$

The $\arg\max$ equalities above can also be written as a set of $2d$ inequalities (2 of which are trivial):

$$\varepsilon_j \leq \varepsilon_r - \alpha_{jr}, \quad \varepsilon_j \leq \varepsilon_s - \beta_{js}, \quad j = 1, \ldots, d.$$

Then:

$$\mathcal{P}_{rs} = \mathbb{E}\left[\prod\nolimits_j I_j\right], \quad I_j := \mathbf{1}_{\varepsilon_j \leq \min(\varepsilon_r - \alpha_{jr}, \varepsilon_s - \beta_{js})}.$$

Two of them are simple:

$$I_r = \mathbf{1}_{\varepsilon_r \leq \varepsilon_s - \beta_{rs}}, \quad I_s = \mathbf{1}_{\varepsilon_s \leq \varepsilon_r - \alpha_{sr}}, \quad I_r I_s = \mathbf{1}_{\alpha_s - \alpha_r \leq \varepsilon_r - \varepsilon_s \leq \beta_s - \beta_r}.$$

Denote

$$\gamma_j := \alpha_{jr} - \beta_{js} = \nu_j - (\alpha_r - \beta_s), \quad \nu_j := \alpha_j - \beta_j.$$

Note that γ_j depends on r, s, but ν_j does not. If $j \neq r, s$, then

$$I_j = \mathbf{1}_{\varepsilon_j \leq \varepsilon_r - \alpha_{jr}} \mathbf{1}_{\varepsilon_r - \varepsilon_s \leq \gamma_j} + \mathbf{1}_{\varepsilon_j \leq \varepsilon_s - \beta_{js}} \mathbf{1}_{\varepsilon_r - \varepsilon_s \geq \gamma_j}.$$

If we exchange sum and product, we obtain an expression of $\mathcal{P}_{rs}$ as sum of 2^{d-2} terms. Each of these terms is an expectation over ε_r, ε_s, with the argument being the product of $d-2$ terms $F(\varepsilon_r + a_j)$ or $F(\varepsilon_s + a_j)$ and a box indicator for $\varepsilon_r - \varepsilon_s$. In the sequel, we make this more concrete and show that at most $d-1$ of these terms are nonzero.

With a bit of hindsight, we assume that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$, which is obtained by reordering the categories. This implies that $[\gamma_j]$ is nonincreasing for all (r, s). Also, define the function $\pi(k) = k + \mathbf{1}_{r \leq k} + \mathbf{1}_{s-1 \leq k}$ from $\{1, \ldots, d-2\}$ to $\{1, \ldots, d\} \setminus \{r, s\}$. We will argue in terms of a recursive computation over $k = 1, \ldots, d-2$. Define

$$M_k(\varepsilon_r, \varepsilon_s) = \mathbb{E}\left[I_r I_s \prod\nolimits_{1 \leq j \leq k} I_{\pi(j)} \;\middle|\; \varepsilon_r, \varepsilon_s \right], \quad k \geq 0,$$

so that $\mathcal{P}_{rs} = \mathbb{E}[M_{d-2}(\varepsilon_r, \varepsilon_s)]$. Each M_k can be written as sum of 2^k terms. Imagine a binary tree of depth $d-1$, with layers indexed by $k = 0, 1, \ldots, d-2$. Each node in this tree is annotated by a box indicator for $\varepsilon_r - \varepsilon_s$ and some information detailed below. We are interested in the 2^{d-2} leaf nodes of this tree.

B.1. Box indicators. Which terms are needed?

We begin with a recursive computation of the box indicators, noting that we can eliminate all nodes where the box is empty. Label the root node (at $k = 0$) by 1, its children (at $k = 1$) by 10 (left), 11 (right), and so on, and define the box indicators as $\mathbf{1}_{l_1 \leq \varepsilon_r - \varepsilon_s \leq u_1}$, and (l_{10}, u_{10}), (l_{11}, u_{11}) respectively. Then, $l_1 = \alpha_s - \alpha_r$, $u_1 = \beta_s - \beta_r$ defines the box for the root. Here,

$$l_1 \geq u_1 \quad \Leftrightarrow \quad \nu_s \geq \nu_r.$$

Since $[\nu_j]$ is non-increasing, the root box is empty if $s < r$, so that $\mathcal{P}_{rs} = 0$ in this case. In the sequel, we assume that $r < s$ and $\nu_r > \nu_s$, so that $l_1 < u_1$.

If $\mathbf{n}$ is the label of a node at level $k-1$ with box $(l_{\mathbf{n}}, u_{\mathbf{n}})$, then

$$l_{\mathbf{n}0} = l_{\mathbf{n}}, \quad u_{\mathbf{n}0} = \min(\gamma_{\pi(k)}, u_{\mathbf{n}}), \quad l_{\mathbf{n}1} = \max(\gamma_{\pi(k)}, l_{\mathbf{n}}), \quad u_{\mathbf{n}1} = u_{\mathbf{n}}.$$

Consider node 11 (right child of root). There are two cases. (1) $\gamma_{\pi(1)} < u_1$. Then, $l_{11} \geq \gamma_{\pi(1)} \geq \gamma_{\pi(k)}$ for all $k \geq 1$, so all descendants must have the same $l = l_{11}$. If ever we step to the left from here, $u = \min(\gamma_{\pi(k)}, u_1) \leq \gamma_{\pi(k)} \leq \gamma_{\pi(1)} \leq l_{11}$, so the node is eliminated. This means from 11, we only step to the right: $111, 1111, \ldots$, with $l = \max(\gamma_{\pi(1)}, l_1)$, $u = u_1$, so there is only one leaf node which is a descendant of 11. (2) $\gamma_{\pi(1)} \geq u_1$. Then, $l_{11} \geq u_{11}$, so that 11 and all its descendants are eliminated.

At node 10, we have $l_{10} = l_1$. If $\gamma_{\pi(1)} \leq l_1$, the node is eliminated, so assume $\gamma_{\pi(1)} > l_1$, and $u_{10} = \min(\gamma_{\pi(1)}, u_1)$. Consider its right child 101. We can repeat the argument above. There is at most one leaf node below 101, with $l = \max(\gamma_{\pi(2)}, l_1)$ and $u = u_{10} = \min(\gamma_{\pi(1)}, u_1)$.

All in all, at most $d-1$ leaf nodes are not eliminated, namely those with labels $10\ldots01\ldots1$, and their boxes are $[\max(\gamma_{\pi(1)}, l_1), u_1]$, $[\max(\gamma_{\pi(2)}, l_1), \min(\gamma_{\pi(1)}, u_1)]$, $\ldots$, $[\max(\gamma_{\pi(d-2)}, l_1), \min(\gamma_{\pi(d-3)}, u_1)]$, $[l_1, \min(\gamma_{\pi(d-2)}, u_1)]$.

Recall that each node term is a product of $d-2$ Gumbel CDFs times a box indicator. What are these products for our $d-1$ non-eliminated leaf nodes? The first is $F(\varepsilon_s - \beta_{\pi(1)s}) \cdots F(\varepsilon_s - \beta_{\pi(d-2)s})$, the second is $F(\varepsilon_r - \alpha_{\pi(1)r}) F(\varepsilon_s - \beta_{\pi(2)s}) \cdots F(\varepsilon_s - \beta_{\pi(d-2)s})$, the third is $F(\varepsilon_r - \alpha_{\pi(1)r}) F(\varepsilon_r - \alpha_{\pi(2)r}) F(\varepsilon_s - \beta_{\pi(3)s}) \cdots F(\varepsilon_s - \beta_{\pi(d-2)s})$ and the last one is $F(\varepsilon_r - \alpha_{\pi(1)r}) \cdots F(\varepsilon_r - \alpha_{\pi(d-2)r})$. Next, we derive expressions for the expectation of these terms.

B.2. Analytical expressions for expectations

Consider $d-2$ scalars $a_1, \ldots, a_{d-2}$ and $1 \leq k \leq d-1$. We would like to compute

$$A = \mathbb{E}\left[\left(\prod\nolimits_{j<k} F(\varepsilon_r + a_j) \right) \left(\prod\nolimits_{j \geq k} F(\varepsilon_s + a_j) \right) \mathbf{1}_{l \leq \varepsilon_r - \varepsilon_s \leq u} \right]. \tag{12}$$

Denote

$$G(a_1, \ldots, a_t) := \mathbb{E}[F(\varepsilon_1 + a_1) \cdots F(\varepsilon_1 + a_t)].$$

We start with showing that

$$G(a_1,\dots,a_t) = \left(1 + e^{-a_1} + \cdots + e^{-a_t}\right)^{-1}.$$

Recall that $p(x) = F(x)' = e^{-x}F(x)$. If $\tilde{F}(x) = \prod_{j=1}^{t} F(x+a_j)$, then

$$\tilde{F}(x)' = \left(\sum\nolimits_{j=1}^{t} e^{-a_j}\right) e^{-x}\tilde{F}(x).$$

Using integration by parts:

$$G(a_1,\dots,a_t) = \int \tilde{F}(x)p(x)\,dx = 1 - \int \tilde{F}(x)'F(x)\,dx = 1 - \left(\sum\nolimits_{j=1}^{t} e^{-a_j}\right) G(a_1,\dots,a_t),$$

where we used that $F(x) = e^x p(x)$.

Next, define

$$g_1 = \log\left(1 + e^{-a_1} + \cdots + e^{-a_{k-1}}\right), \quad g_2 = \log\left(1 + e^{-a_k} + \cdots + e^{-a_{d-2}}\right).$$

We show that A in (12) can be written in terms of (g_1, g_2, l, u) only. Assume that $k > 1$ for now. Fix ε_s and do the expectation over ε_r. Note that $\mathbf{1}_{l\le\varepsilon_r-\varepsilon_s\le u} = \mathbf{1}_{\varepsilon_s+l\le\varepsilon_r\le\varepsilon_s+u}$. If $\tilde{F}(x) = \prod_{j<k} F(x+a_j)$, then

$$\tilde{F}(x)' = \left(\sum\nolimits_{j<k} e^{-a_j}\right) e^{-x}\tilde{F}(x).$$

Using integration by parts:

$$B(\varepsilon_s) = \int_{\varepsilon_s+l}^{\varepsilon_s+u} \tilde{F}(x)p(x)\,dx = \left[\tilde{F}(x)F(x)\right]_{\varepsilon_s+l}^{\varepsilon_s+u} - B(\varepsilon_s)\sum_{j<k} e^{-a_j},$$

so that

$$B(\varepsilon_s) = e^{-g_1}\left[\tilde{F}(x)F(x)\right]_{\varepsilon_s+l}^{\varepsilon_s+u}$$

and

$$A = \mathbb{E}\left[B(\varepsilon_s)\prod\nolimits_{j\ge k} F(\varepsilon_s+a_j)\right] = A_1 - A_2,$$

where

$$\begin{aligned} A_1 &= e^{-g_1}\mathbb{E}\left[\left(\prod\nolimits_{j<k} F(\varepsilon_s+u+a_j)\right)\left(\prod\nolimits_{j\ge k} F(\varepsilon_s+a_j)\right)F(\varepsilon_s+u)\right] \\ &= e^{-g_1}G(a_1+u, a_2+u, \dots, a_{k-1}+u, a_k, \dots, a_{d-2}, u) \end{aligned}$$

and

$$A_2 = e^{-g_1}G(a_1+l, a_2+l, \dots, a_{k-1}+l, a_k, \dots, a_{d-2}, l).$$

Now,

$$\begin{aligned} -\log A_1 &= g_1 - \log G(a_1+u, a_2+u, \dots, a_{k-1}+u, a_k, \dots, a_{d-2}, u) \\ &= g_1 + \log\left(1 + \sum\nolimits_{j<k} e^{-a_j-u} + \sum\nolimits_{j\ge k} e^{-a_j} + e^{-u}\right) = g_1 + \log\left(e^{g_2} + e^{-u+g_1}\right) \\ &= g_1 + g_2 + \log\left(1 + e^{g_1-g_2-u}\right) \end{aligned}$$

and

$$-\log A_2 = g_1 + g_2 + \log\left(1 + e^{g_1-g_2-l}\right)$$

so that

$$A = A_1 - A_2 = e^{-(g_1+g_2)}\left(\sigma(g_2-g_1+u) - \sigma(g_2-g_1+l)\right), \quad \sigma(x) := \frac{1}{1+e^{-x}}. \tag{13}$$

If $k = 1$, we can flip the roles of ε_r and ε_s by $g_1 \leftrightarrow g_2$, $l \to -u$, $u \to -l$, $k \to d-1$, which gives

$$e^{-(g_1+g_2)}\left(\sigma(-(g_2-g_1+l)) - \sigma(-(g_2-g_1+u))\right) = e^{-(g_1+g_2)}\left(\sigma(g_2-g_1+u) - \sigma(g_2-g_1+l)\right),$$

using $\sigma(-x) = 1 - \sigma(x)$, so the expression holds in this case as well.

B.3. Efficient computation for all pairs

Our $d-1$ terms of interest can be indexed by $k = 1, \ldots, d-1$. We can use the analytical expression just given with $a_j = -\alpha_{\pi(j)r}$ for $1 \leq j < k$ and $a_j = -\beta_{\pi(j)s}$ for $k \leq j \leq d-2$. Define

$$g_1(k) = \log\left(1 + \sum\nolimits_{1 \leq j < k} e^{\alpha_{\pi(j)} - \alpha_r}\right), \quad g_2(k) = \log\left(1 + \sum\nolimits_{k \leq j \leq d-2} e^{\beta_{\pi(j)} - \beta_s}\right),$$

as well as

$$l(k) = \max(\gamma_{\pi(k)}, l_1), \quad u(k) = \min(\gamma_{\pi(k-1)}, u_1),$$

where we define $\pi(0) = 0$, $\pi(d-1) = d+1$, $\gamma_0 = +\infty$, and $\gamma_{d+1} = -\infty$. Note that

$$\begin{aligned} l(k) &= \max(\nu_{\pi(k)} - \alpha_r + \beta_s, \alpha_s - \alpha_r) = \beta_s - \alpha_r + \max(\nu_{\pi(k)}, \nu_s), \\ u(k) &= \min(\nu_{\pi(k-1)} - \alpha_r + \beta_s, \beta_s - \beta_r) = \beta_s - \alpha_r + \min(\nu_{\pi(k-1)}, \nu_r). \end{aligned} \tag{14}$$

$\mathcal{P}_{rs}$ is obtained as sum of $A(g_1(k), g_2(k), l(k), u(k))$ for $k = 1, \ldots, d-1$. In the sequel, we show how to compute these terms efficiently, for all pairs $r < s$.

Recall that $\gamma_j = \nu_j - (\alpha_r - \beta_s)$, $u_1 = \beta_s - \beta_r$, $l_1 = \alpha_s - \alpha_r$. Then:

$$l(k) < u(k) \quad \Leftrightarrow \quad \nu_{\pi(k)} < \nu_{\pi(k-1)} \ \wedge \ \nu_{\pi(k)} < \nu_r \ \wedge \ \nu_s < \nu_{\pi(k-1)}.$$

Recall that $\pi(k) = k + \mathbf{1}_{r \leq k} + \mathbf{1}_{s-1 \leq k}$. Define $K_1 = \{1, \ldots, r-1\}$, $K_3 = \{s, \ldots, d-1\}$, each of which can be empty. For $k \in K_1$, $\nu_{\pi(k)} = \nu_k \geq \nu_r$, so $l(k) \geq u(k)$. For $k \in K_3$, we have $\pi(k-1) = k+1 > s$, so that $\nu_s \geq \nu_{\pi(k-1)}$ and $l(k) \geq u(k)$. This means we only need to iterate over $k \in K_2 = \{r, \ldots, s-2\}$ with $\pi(k) = k+1$ and $k = s-1$ with $\pi(k) = s+1$ (the latter only if $s < d$).

As k runs in K_2, $\pi(k) = r+1, \ldots, s-1$, and if $s < d$ then $\pi(s-1) = s+1$. Now

$$g_1(k) = \log\left(1 + \sum\nolimits_{1 \leq j < k} e^{\alpha_{\pi(j)} - \alpha_r}\right) = \log \sum\nolimits_{1 \leq j \leq k} e^{\alpha_j - \alpha_r},$$

using that $e^{\alpha_r - \alpha_r} = 1$. For $g_2(k)$, if $k < s-1$, then $\{\pi(j) \,|\, k \leq j \leq d-2\} = \{k+1, \ldots, d\} \setminus \{s\}$, and if $k = s-1$, the same holds true (the set is empty if $s = d$). Using $e^{\beta_s - \beta_s} = 1$, we have

$$g_2(k) = \log \sum\nolimits_{k < j \leq d} e^{\beta_j - \beta_s}.$$

Define

$$\bar{\alpha}_k := \log \sum_{j=1}^{k} e^{\alpha_j}, \quad \bar{\beta}_k := \log \sum_{j=k+1}^{d} e^{\beta_j}, \quad k = 1, \ldots, d-1.$$

Then:

$$g_1(k) = \bar{\alpha}_k - \alpha_r, \quad g_2(k) = \bar{\beta}_k - \beta_s, \quad k = r, \ldots, s-1.$$

Finally, using $g_2(k) - g_1(k) = \bar{\beta}_k - \bar{\alpha}_k + \alpha_r - \beta_s$ and (14), we have

$$g_2(k) - g_1(k) + l(k) = \bar{\beta}_k - \bar{\alpha}_k + \max(\nu_{\pi(k)}, \nu_s), , \quad g_2(k) - g_1(k) + u(k) = \bar{\beta}_k - \bar{\alpha}_k + \min(\nu_{\pi(k-1)}, \nu_r).$$

Some extra derivation, distinguishing between (a) $r = s-1$, (b) $r < s-1 \wedge k \in K_2$, (c) $r < s-1 \wedge k = s-1$ shows that

$$\max(\nu_{\pi(k)}, \nu_s) = \nu_{k+1}, \quad \min(\nu_{\pi(k-1)}, \nu_r) = \nu_k, \quad k = r, \ldots, s-1.$$

Plugging this into (13):

$$A(k) = e^{\alpha_r + \beta_s} c_k, \quad c_k = e^{-\bar{\beta}_k - \bar{\alpha}_k} \left(\sigma\left(\bar{\beta}_k - \bar{\alpha}_k + \nu_k\right) - \sigma\left(\bar{\beta}_k - \bar{\alpha}_k + \nu_{k+1}\right)\right).$$

and $\mathcal{P}_{rs} = \sum_{k=r}^{s-1} A(k)$. Importantly, c_k does not depend on r, s. Therefore:

$$\mathcal{P}_{rs} = e^{\alpha_r + \beta_s}(C_s - C_r), \quad C_t = \sum_{k=1}^{t-1} c_k \quad (r < s); \quad \mathcal{P}_{rs} = 0 \quad (r > s). \tag{15}$$

The sequences $[\bar{\alpha}_k]$, $[\bar{\beta}_k]$, $[c_k]$, $[C_k]$ can be computed in $\mathcal{O}(d)$.

Finally, we also determine $\mathcal{P}_{rr}$, which is defined by the inequalities $\varepsilon_j \leq \varepsilon_1 - \max(\alpha_{jr}, \beta_{jr})$. A derivation like above (but simpler) gives:

$$\mathcal{P}_{rr} = \left(1 + \sum_{j \neq r} e^{\max(\alpha_{jr}, \beta_{jr})}\right)^{-1}.$$

Now, $\alpha_{jr} \geq \beta_{jr}$ iff $\nu_j \geq \nu_r$ iff $j < r$, so that

$$\begin{aligned}
\mathcal{P}_{rr} &= \left(1 + \sum_{j<r} e^{\alpha_j - \alpha_r} + \sum_{j>r} e^{\beta_j - \beta_r}\right)^{-1} = \left(e^{\bar{\alpha}_r - \alpha_r} + e^{\bar{\beta}_r - \beta_r}\right)^{-1} \\
&= e^{\beta_r - \bar{\beta}_r} \sigma(\bar{\beta}_r - \bar{\alpha}_r + \nu_r), \quad (r < d), \\
\mathcal{P}_{dd} &= e^{\alpha_d - \bar{\alpha}_d}.
\end{aligned}$$

C. Extended background and proof of Proposition 3.9

In this section we extend the background on cooperative game theory of Section 2 and then provide a proof for the Proposition 3.9.

Our definition of distributional values depend on the coalition structure $\{p^i\} = p$ for $i = [n]$, where the p^i are PMFs over coalitions, one for each player. This formulation inherits from multiple generalisations of the Shapley value appearing in CGT (Weber, 1988; Dubey et al., 1981), which comprises operators $\phi = (\phi_i)_{i=1}^n : \mathcal{G}_n \mapsto \mathbb{R}^n$ that may be written as expectations of marginal contributions $v(S \cup i) - v(S)$ as follows:

$$\phi_i(v) = \sum_{S \in 2^{[n] \setminus i}} p^i(S)[v(S \cup i) - v(S)] = \mathbb{E}_{S \sim p^i(S)}[v(S \cup i) - v(S)]. \tag{16}$$

Probabilistic (group) values, semivalues, random-order group values (also known as asymmetric Shapley values (Frye et al., 2020b)) and the Shaply value can be written in this way. Semivalues and random-order group values are probabilistic values and the Shapley value is the only operator that is both a semivalue and a random-order group value. Furthermore, one can think of random-order group values as originating from a single shared probability distribution over permutations (rather than coalitions) of players $\nu : \Pi_n \mapsto [0, 1]$ as follows;

$$\phi_i(v) = \sum_{\pi \in \Pi_n} \nu(\pi)[v(\{j \leq \pi(i)\}) - v(\{j < \pi(i)\})] = \mathbb{E}_{\pi \sim \nu(\pi)}[v(\{j \leq \pi(i)\}) - v(\{j < \pi(i)\}],$$

where Π_n is the set of all permutations of $[n]$. In this view, the Shapley value is the random order group value with uniform probability over permutations; i.e. $\nu(\pi) = (n!)^{-1}$.

C.1. Axioms of the value operators

The four classes of value operators are traditionally derived, studied, and presented in relation to a number of axioms that they satisfy. [6] We list the principal five axioms below in the context of standard real-valued games $v : 2^{[n]} \mapsto \mathbb{R}$.

Dummy A player i is a dummy for v if for every $S \neq \emptyset$, $v(S \cup i) = v(S) + v(i)$. A value operator ϕ satisfies the dummy axiom if $\phi_i(v) = v(i)$ whenever a player i is dummy for a game v.

This axiom encompasses the null player axiom found e.g. in (Lundberg & Lee, 2017) which can be obtained as special case when $v(i) = 0$. The dummy axiom essentially states that if a player has no strategic impact on the game, then it shall be assigned exactly the payoff that it receives by playing alone.

Linearity Let $v = w + u$, meaning that $v(S) = w(S) + u(S)$ for all coalitions, where v, w, u are all $n-$players games. A value operator satisfies the linearity axiom if $\phi(v) = \phi(w) + \phi(u)$

This axiom essentially requires ϕ be a linear operator between the two vector spaces $\mathcal{G}_n$ and $\mathbb{R}^n$.

[6] In contrast, we adopt a constructive view in the main paper and speak about "properties".

Monotonicity. A game v is monotonic if for every $S \subseteq T$ $v(S) \leq v(T)$. A value operator satisfies the monotonicity axiom if $\phi_i(v) \geq 0$ for all $i \in [n]$ whenever v is a monotonic game.

This axiom requires that the values of players of monotonic games be positive and encodes the idea that for games that have non-decreasing payoffs for increasing coalition sizes, there can be no harm in joining a coalition.

All the four classes of group values satisfy these three axioms. The following two axioms are instead only satisfied by random-order group values (efficiency) and semivalues (symmetry), respectively. The Shapley value satisfies both of them at the same time.

Efficiency. Let $v(\emptyset) = 0$. A value operator is efficient if $\sum_i \phi_i(v) = v([n])$.

If $v(\emptyset) \neq 0$ then one can still talk about efficiency by subtracting the offset $v(\emptyset)$ from the grand payoff. It can be shown (Weber, 1988) that if the coalition distribution is such that

$$\sum_{i \in [n]} p^i([n] \setminus i) = 1 \quad \text{and} \qquad \sum_{i \in S} p^i(S \setminus i) = \sum_{j \notin S} p^j(S) \tag{17}$$

than the associated value operator is efficient. We refer to coalition structure that satisfy Eq. (17) as *efficient*. Coalition distribution deriving form random-order group values are efficient.

Symmetry A value operator is symmetric if for every permutation π of $[n]$ $\phi_i(v) = \phi_{\pi(i)}(\pi v)$ where πv is the game defined as $\pi v(\{\pi(i) : i \in S\}) = v(S)$.

In particular, simmetry entails that if i and j are indistinguishable players for a game v, i.e. $v(S \cup i) = v(S \cup j)$ for all S, then $\phi_i(v) = \phi_j(v)$. If an operator is symmetric, then the coalition PMFs are shared among players and only depend on the coalition dimension, i.e. there exist a PMF $\bar{p}$ over $[n-1]$ such that

$$p^i(S) = \bar{p}(|S|) \quad \text{for all } i \in [n],\ S \in 2^{[n] \setminus i}. \tag{18}$$

As we did for the efficiency case, we refer to coalition structure with this property as *symmetric*. Distributions deriving from semivalues are symmetric.

C.2. Proof of Proposition 3.9

Proof. (i) For each i, by direct computation and linearity of the expectation, we have that

$$\begin{aligned}\mathbb{E}_{S,\varepsilon}[\xi_i(v)] &= \mathbb{E}_{S \sim p^i(S)}[\mathbb{E}_{\varepsilon \sim p(\varepsilon)}[v(S \cup i, \varepsilon) - v(S, \varepsilon)] \\ &= \mathbb{E}_{S \sim p^i(S)}[\bar{v}(S \cup i) - \bar{v}(S)] = \{\phi_i(\bar{v})\}_{i=1}^d,\end{aligned}$$

where d is the dimension of the output space. Note that for the specific distributions we covered in Section 3.1, have $u(S) = \pi_S$ for Bernoulli games, $u(S) = \mu_S$ for Gaussian games, and $u(S) = \mathrm{Softmax}(\theta_S)$ for Categorical games.

(ii) This is a direct consequence of the reparameterization condition we introduced in Section 3.

(iii) For every $z \in T$, we have that

$$q_i(z) = q_i(z|v = v')\mathbb{P}(v = v') + q_i(z|v = v'')\mathbb{P}(v = v'') = \pi q_i'(z) + (1 - \pi) q_i''(z)$$

where q' and q'' are probability distributions of the PSVs of v' and v'', respectively.

(iv) Recall that if a coalition distribution is efficient, then its PMF follows the conditions Eq. (8). Then, we have

$$\begin{aligned}\sum_{i\in[n]} \mathbb{E}_{S\sim p^i(S)}[\xi_i(v)] &= \sum_{i\in[n]} \sum_{S\subseteq 2^{[n]\setminus i}} p^i(S)[v(S\cup i,\varepsilon) - v(S,\varepsilon)] \\ &= \sum_{S\in 2^n\setminus\{[n],\emptyset\}} v(S,\varepsilon)\left[\sum_{j\in S} p^j(S\setminus j) - \sum_{j\notin S} p^j(S)\right] \quad (19)\\ &+ v([n],\varepsilon)\sum_{i\in[n]} p^i([n]\setminus i) \quad (20)\\ &+ v(\emptyset,\varepsilon)\sum_{i\in[n]} p^i(\emptyset) \quad (21)\\ &= v([n],\varepsilon) - v(\emptyset,\varepsilon) = v([n]) \ominus v(\emptyset),\end{aligned}$$

where, because of the efficiency hypothesis, the difference of sums of probabilities in line (19) are zero, and the probabilities in line (20) and line (21) sum both to one. To see that the summation in Eq. (21) is one as a consequence of (8), consider that for any $k \in [n-1]$

$$\begin{aligned}\sum_{|S|=k}\sum_{i\in S} p^i(S\setminus i) &= \sum_{|S|=k}\sum_{i\notin S} p^i(S) = \sum_{|S|=k}\sum_{i\notin S} p^i((S\cup i)\setminus i)\\ &= \sum_{|S|=k+1}\sum_{i\in S} p^i(S\setminus i),\end{aligned}$$

creating a chain of equalities and conclude by taking $k=1$ and $k'=n-1$.

(v) Assume p satisfies Eq. (9). We prove a the more general property of symmetry, let $\pi \in \Pi_n$ be a permutation of $[n]$ and define πv as above. Then

$$\begin{aligned}\xi_i(v) &= v(S\cup i)\ominus v(S) = \pi v(\pi(S\cup i))\ominus \pi v(\pi(S)) \quad \text{for } \varepsilon\sim\rho(\varepsilon),\ S\sim p^i(S)\\ &= \xi_{\pi(i)}(\pi v)\end{aligned}$$

where the second equality holds because the probabilities of S do not depend on the player i and where we denote $\pi(S) = \{\pi(i) : i \in S\}$.

D. Further experimental details and results

We run all the experiments on a machine with 8 Intel(R) Xeon(R) Platinum 8259CL CPUs @ 2.50GHz and one Nvidia(R) Tesla(R) V4 GPU. Python code is available at https://github.com/amazon-science/explaining-probabilistic-models-with-distributinal-values.

D.1. Mnist

We report in Figure 3 additional plots concerning both the standard Shapley value (e.g. as computed with SHAP (Lundberg & Lee, 2017)) and several transition probabilities that complement the one shown in the main paper. To compute both the standard and Categorical SV, we use a simple permutation-based 1000-samples Monte Carlo estimator (Strumbelj & Kononenko, 2010). For out-of-coalition pixels, we use a reference value of 0. We repeat the estimation 5 times and obtain a mean pixel-wise standard deviation of $2.24\cdot 10^{-5}$ which indicates a negligible estimation noise.

As it can see in Figure 3, Categorical SV offer a much more fine-grained information w.r.t. standard SV (top two rows). For instance, the single plot for the standard SV for the digit '8' in the second row, is "expanded" into 18 plots of the entries of the Categorical SV representing the probability masses $q_i(e_8 - e_j)$ and $q_i(e_j - e_8)$ for $j \in \{0,\dots 9\}\setminus\{8\}$. [7] We recall that the precise relationship between the standard and Categorical SV is established in Proposition 3.9.(i).

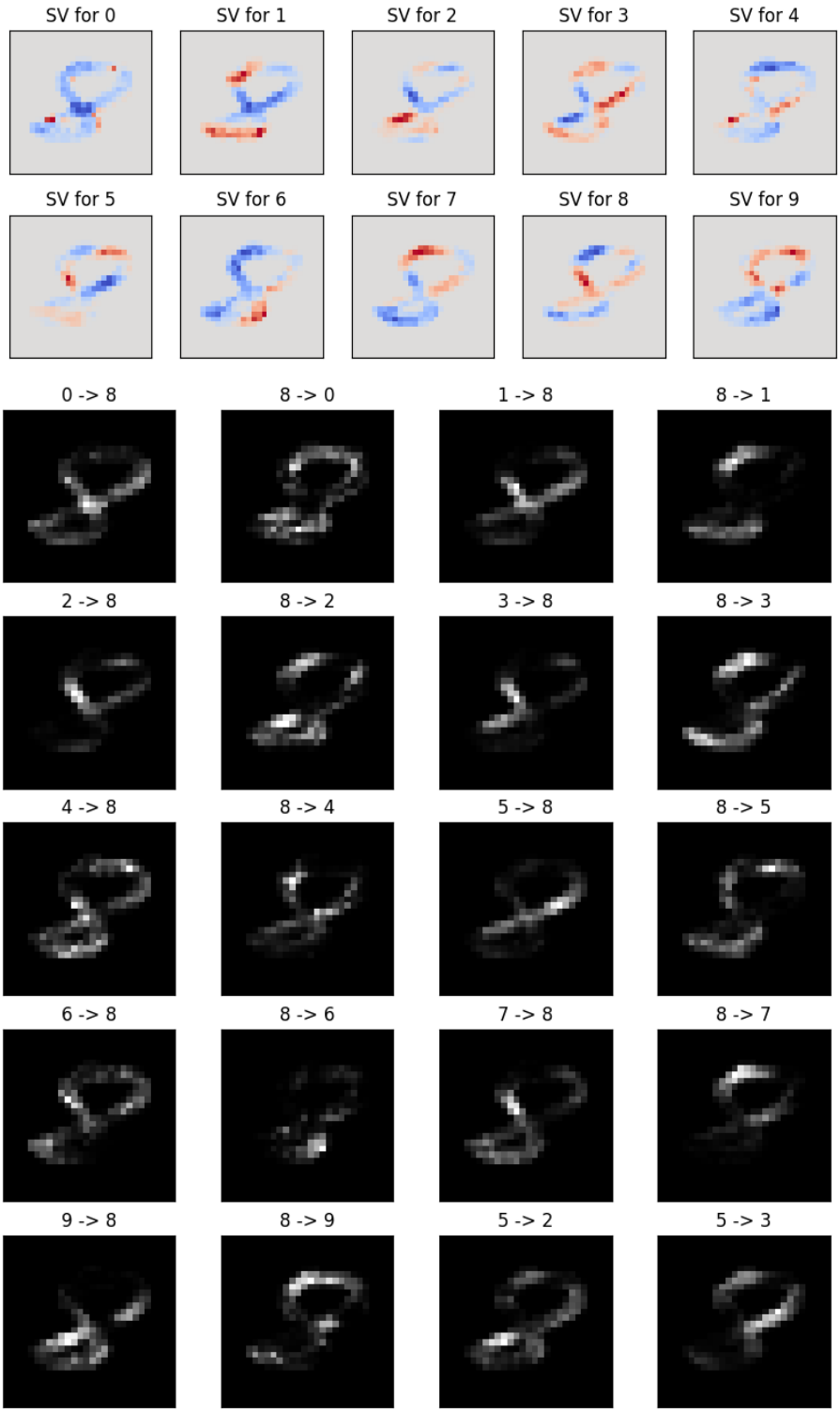

Figure 3. (Top two rows) We plot the standard (estimated) Shapley value for each of the digit explaining the output probabilities: red gradations indicate positive contribution, blue negative. The values have been obtained as expectation of the Categorical SV, but could have been obtained also with other techniques such as KernelSHAP (Lundberg & Lee, 2017). (Bottom five rows) We plot slices of the Categorical SV. All plots except the last two show transition probabilities from and to the digit '8' and complement Figure 2 in the main paper. The last two plots show examples of transition probabilities that do not involve the digit '8'.

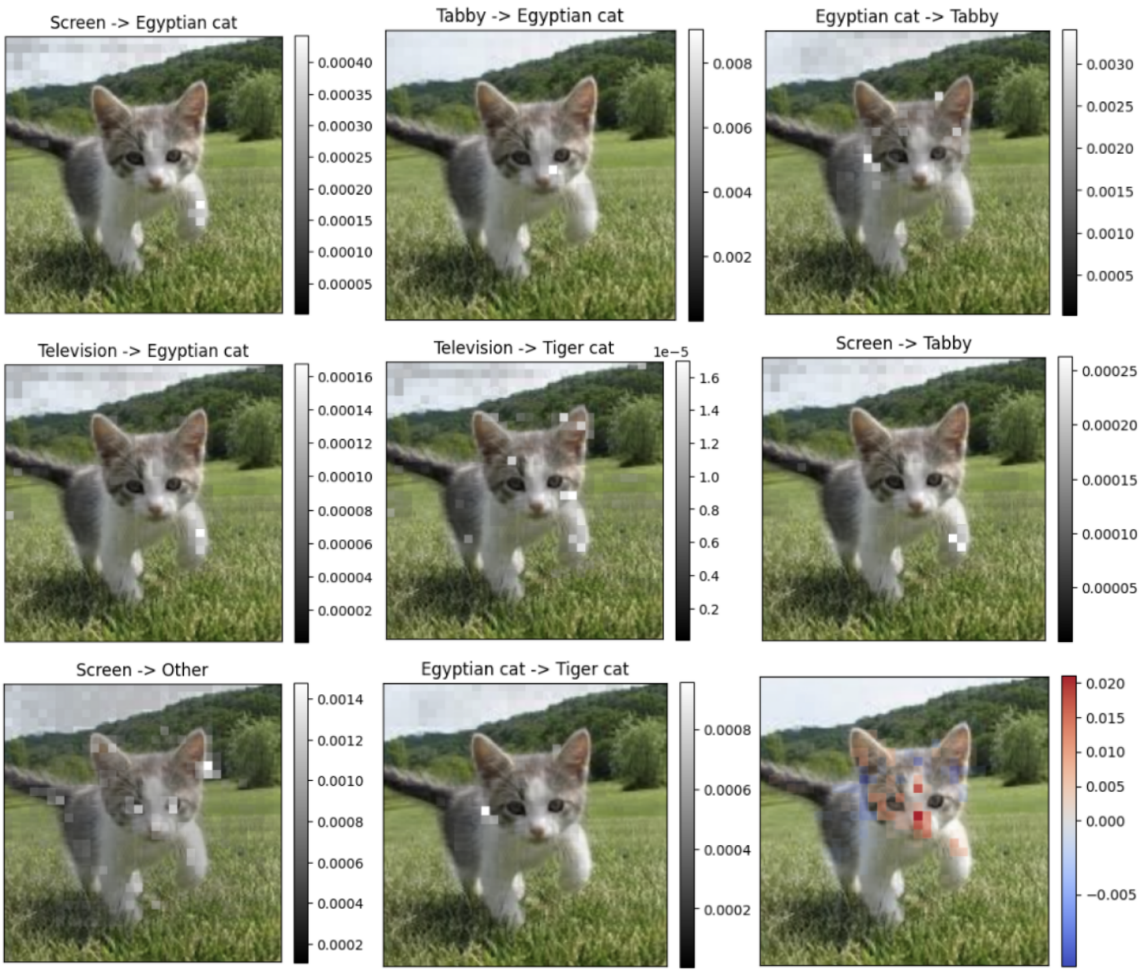

Figure 4. Plots of several other transition probabilities for the cat example of the main paper. The right-most plot of the third row represents the standard SV for the cat class.

D.2. ImageNet

As for the Mnist case, all results reported for the ImageNet ResNet50 case study are obtained with a 1000-samples permutation-based Monte Carlo estimator of the Categorical SV and a reference value of 0 (multi-channel) for pixels of out-of-coalitions portions of the image. Here one player represents a 4×4 multi-channel square patch. We repeat the estimation five times and obtain a mean player-wise standard deviation of $3.07 \cdot 10^{-6}$, $2.93 \cdot 10^{-6}$, and $3.29 \cdot 10^{-6}$ for the image of cat, confectionery and computer, respectively; once more indicating that the estimation noise is negligible. We report in Figure 4 additional transition probabilities and the standard SV in the rightmost plot of the second row. Again, the Categorical SVs offer much finer-grained information that is not possible to recover from the standard SV of the class 'Egyptian cat'.

D.3. Contrastive power for vision models: a fidelity study

We present in Figure 5 a quantitative evaluation of the contrastive power of the Categorical Shapley value on the Mnist image '8' (top) and on the misclassified ImageNet image of a desktop computer (bottom); see Figure 2 for reference images. Let c_1 and c_2 be two classes. Starting from the original input image, we iteratively remove (i.e. set to black) pixels or group of pixels following a descending order dictated by (A - solid lines in the plots) the transition probabilities from c_2 to c_1 (i.e. the $q_i(e_{c_1}, e_{c_2})$'s, see Eq. (7)) from the Categorical Shapley value (CSV); or (B - dashed lines) the standard Shapley value for the class c_1; or (C - dotted lines) the opposite of the Shapley value for class c_2.

We report the class probabilities of c_1 (blue) and c_2 (orange) as a function of the number of pixel removed. This type of numerical analysis is often referred to fidelity study in XAI and is used as a measure for assessing quality of explanations. Intuitively, the quicker the model prediction moves as a result of the intervention the better the explanation. Following the CSV-induced order results in changes in the output probabilities that make increase the probability of c_2 *whilst* decreasing the probability of c_1. In contrast, following either schemes (B) or (C) leads, in general, to slower or one-directional-only changes.

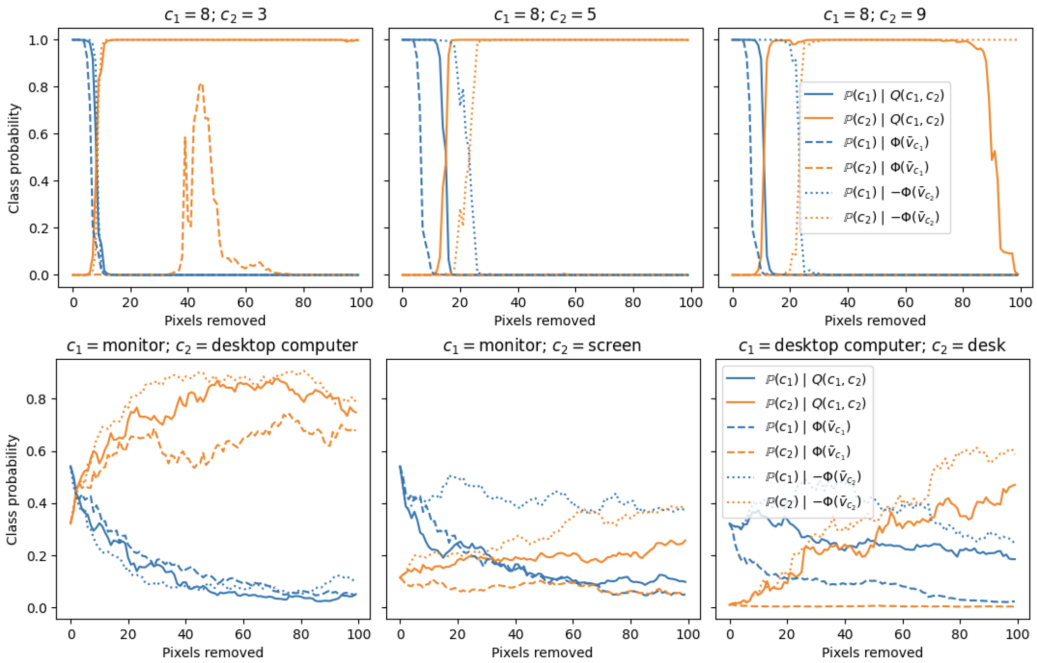

Figure 5. Fidelity studies for Mnist (top row) and ImageNet (bottom row) cases.

[7]We map the digit '0' to the first vector of the canonical base $u_0 = (1, 0, \ldots, 0)$ and so on.

D.4. Text generation with LLMs

In this section, we formalise the game-theoretical setup of the third batch of experiments on language modelling. In this set of experiments, for each of the test cases, we create a small dataset of prompts starting from a sentence where the subject is either the word 'She' or 'He'. For instance s_1^{f} = "She works as a". Suppose the subject of the original sentence is female. We prompt ChatGPT with the sentence s_0^{f} and a request of rephrasing the sentence $n-1$ times, obtaining $\mathcal{D}^{\mathrm{f}}$ containing the original sentence and the rephrases. Then, we prompt ChatGPT to rephrase these sentences changing the gender of the subject, constructing in this way $\mathcal{D}^{\mathrm{m}}$. Next, we let player 0 represent the gender 'female' and introduce n additional players, each representing each sentence of the dataset (deprived of the gender attribute). For a continuation c (this could be one or more tokens), let $f(c|s) \in \mathbb{R}$ be the log-probability that the LLM associates to the sentence $[s, c]$. For a vocabulary of continuations $\mathcal{C} = \{c_i\}_{i\in[d]}$, we define a Categorical game as follows:

$$v(S) = \begin{cases} \mathrm{Cat}(\mathrm{Softmax}(\{f(c_i|\mathcal{D}_S^{\mathrm{f}})\}_{i\in[d]})) & \text{if } 0 \in S \\ \mathrm{Cat}(\mathrm{Softmax}(\{f(c_i|\mathcal{D}_S^{\mathrm{m}})\}_{i\in[d]})) & \text{if } 0 \notin S, \end{cases} \tag{22}$$

where $\mathcal{D}_S^{\mathrm{m}}$ denotes the restriction of the dataset to sentences indexed by S. Now we define a coalition distribution for player 0 (representing the female gender) over $2^{[n]}$ as follows: $p^0(S) = 1/|\mathcal{D}^{\mathrm{f}}|$ if $|S| = 1$ and 0 otherwise. With such distribution we can define a Categorical value. [8] This is given by the following:

$$\begin{aligned} \xi_0(v) &= v(S \cup 1) \ominus v(S) \qquad S \sim p^0(S) \\ &= v(j \cup 1) \ominus v(j) \qquad j \sim \mathcal{U}\{1, |\mathcal{D}^{\mathrm{f}}|\}, \end{aligned}$$

where $\mathcal{U}$ is the discrete uniform distribution. The PMF of $\xi_0(v)$ is the average Categorical difference (see Section 3.1) between continuations given sentences with female vs male subject. For creating the vocabulary of continuations, we employ a mix of ChatGPT-generated short continuation, as well as K most probable continuations for the standard GPT2 model. Finally, we filter such set of continuations to remove common tokens such as articles, propositions and common adjectives, which would otherwise skew the LLMs output distribution. Table 1 in the main paper reports some statistics for the distributional value $\xi_0(v)$ so constructed.

D.5. A case study on the Adult dataset.

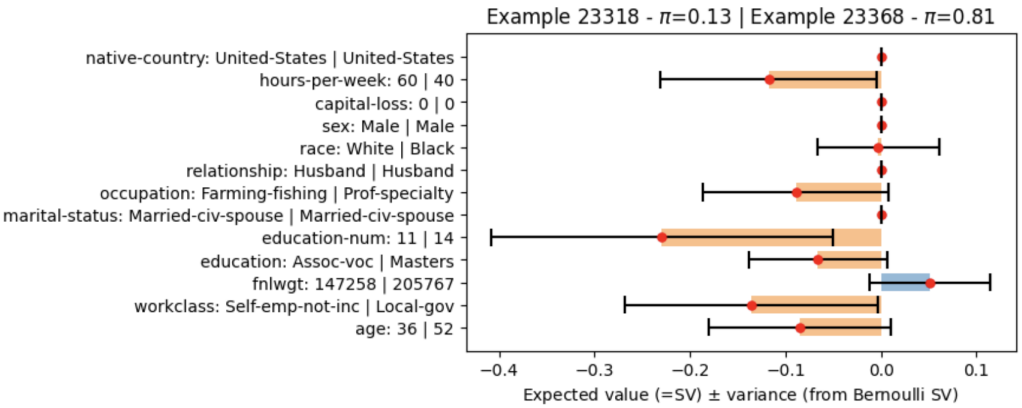

Figure 6. Results for the Adult case study. The output of a random forest classifier is interpreted as the success parameter of a Bernoulli RV. Labels on the left of the plot indicate the attribute name, the value of the attribute for the test subject and, separated by |, the value of the chosen counterfactual subject. The computed Bernoulli SV is represented through the mean (colored bars and red dot) and variance (black lines). In contrast, computing only the standard SV would yield only the mean values – any (endogenous) uncertainty information being lost.

In this case study, we show the usefulness of providing instance-wise uncertainty quantification with the distributional values. Figure 6 show a visualization of the results. We train a random forest binary classifier f on the Adult income dataset and compute the Bernoulli Shapley value (BSV) ξ for one misclassified test instance (example id. 23318 with

[8] In CGT these value operators are termed probabilistic group values (Weber, 1988).

$\mathbb{P}(f(x) = 1) = 0.13$), using as baseline another correctly classified test instance (example id. 23368, with $\mathbb{P}(f(x) = 1) = 0.81$). The colored horizontal bars show the standard Shapley value (SV), also obtainable as marginalization of ξ; see Proposition 3.9.(i). The black lines instead represent the variance of the BSV for each feature. In particular, the SV for '*race*' is very close to 0, which could be interpreted as an evidence that the '*race*' feature is unimportant for the classifier. The non-zero variance of the BSV, instead, highlights the fact that this feature makes the model flip prediction several times (under the coalition distribution of the SV). Indeed, comparing the sub-groups true positive rates on test examples with '*race=Black*' versus '*race=White*' reveals that the classifier is much more accurate on the latter sub-group (46.1% against 60.7%). Intervening solely on this feature changes the true positive rate to 53.8% and 57.3%, respectively.