Stability of game-theoretic feature explanations for machine learning models

Alexey Miroshnikov*, Konstandinos Kotsiopoulos†, Khashayar Filom†, and Arjun Ravi Kannan†

Abstract. In this article, we study feature attributions of machine learning models originating from linear game values defined as operators on appropriate functional spaces. The main focus is on random games based on the conditional and marginal expectations. It is well-known from the Rashomon effect that, under predictor dependencies, distinct models that approximate the same data well can have different representations. To understand the impact of the Rashomon effect on the explanation maps, we formulate a stability theory for these explanation operators by establishing certain bounds for both marginal and conditional explanations. The differences between the two games are then elucidated. Specifically, we encode the feature dependencies by the Radon-Nikodym derivative (when it exists) of the marginal probability measure $\tilde{P}_X := \frac{1}{2^n}\sum_{S\subseteq N} P_{X_S}\otimes P_{X_{-S}}$ with respect to the joint measure P_X, and show that the marginal explanation map can become discontinuous on $L^2(P_X)$ as the strength of dependencies in features increases, while the conditional explanations remain stable. We also establish bounds on the distance between marginal and conditional maps in terms of the dependencies. Our analysis illustrates that when dependencies are strong, the marginal explanations do not serve as good approximates of conditional explanations.

Key words. ML interpretability, explanation operator, game value, Radon-Nikodym derivative.

MSC codes. 91A06, 91A12, 91A80, 46N30, 46N99, 68T01

1. Introduction.

The use of Machine Learning (ML) models has become widespread due to their superior performance over traditional statistical techniques. In particular, contemporary ML models have a complex structure which allows for a higher predictive power and the capability of processing a larger number of attributes. Having a complex model structure, however, comes at the expense of increased difficulty in interpretability[1]. This, in turn, may raise concerns of model trustworthiness and create other issues if not appropriately managed.

Explaining the outputs of complex ML models (such as ensemble trees or neural nets) has applications in several fields. Predictive models, and strategies that rely on such models, are sometimes subject to laws and regulations, such as the Equal Credit Opportunity Act which requires financial institutions to notify consumers who have been declined or negatively impacted by a credit decision of the main factors that contributed to that decision. Another application is in medicine, where ML models are used to predict the likelihood of a certain disease or a medical condition, or the result of a medical treatment [22, 46]. Model interpretations (or explanations) can then be used to make judgments regarding the most contributing factors affecting the likelihood of the disease or the choice of the most optimal treatment; for

*Emerging Capabilities Research Group, Discover Financial Services Inc., Riverwoods, IL 60015, USA.
Email: amiroshn@terpmail.umd.edu (first & corresponding author), ORCID:0000-0003-2669-6336,
kkotsiop@gmail.com, ORCID:0000-0003-2651-0087,
khashayar.1367@gmail.com, ORCID:0000-0002-6881-4460,
arjun.kannan@gmail.com, ORCID:0000-0003-4498-1800.

[1]We use the words interpretability (interpretation) and explainability (explanation) interchangeably. However, the methods discussed in this paper primarily deal with post-hoc explanations derived from model's results; for details on interpretable models vs post-hoc explanations see [18].

instance, see [13].

The objective of a model explainer is to quantify the contribution of each predictor to the value of a predictive model f trained on the data (X, Y), where $X \in \mathbb{R}^n$ are predictors and Y is a response variable. Many post-hoc explanations (in the ambient settings) are based on the pair (X, f). However, there are numerous methods that rely on the structure of the model, its implementation, and even the sequence of algorithmic steps that led to the construction of such a model.

There is a comprehensive body of research that discusses approaches for constructing post-hoc explainers as well as self-explainable models. When it comes to explanations, there are global methods such as [17, 28] which quantify the overall effect of features, as well as local methods such as the rule-based method [39], locally-interpretable methods [38, 20], and methods such as [43, 30, 9] which provide individualized feature attributions (for a single prediction) based on the game-theoretic work of Shapley [42]. See also works on explainable neural networks [47] and self-explainable models [3, 14], among others.

Many promising interpretability techniques utilize ideas from the cooperative game theory for constructing explainers using game values with appropriately designed games adopted to a machine learning setting [43, 29, 48, 9, 33, 11, 46, 41, 10, 15]. In this setting, given a model f, the features $X = (X_1, \ldots, X_n)$ are viewed as n players playing a random cooperative game $v(S; X, f)$, a set function on the subsets of indices $S \subseteq \{1, 2, \ldots, n\}$ whose values are random variables where the randomness comes from the features. While the literature considers deterministic games (that are observations of random games), in our paper, we will view them as random in order to perform rigorous analysis.

The two most notable games based on (X, f) are the conditional and marginal games[2]

$$v^{CE}(S) := \mathbb{E}[f(X)|X_S], \quad v^{ME}(S) := \mathbb{E}[f(x_S, X_{-S})]\big|_{x_S = X_S}, \quad S \subseteq N := \{1, 2, \ldots, n\}$$

where $X_S := (X_{i_1}, \ldots, X_{i_k})$, $S = \{i_1, \ldots, i_k\}$, and $-S := N \setminus S$. These are motivated by the corresponding deterministic games introduced in [43, 30] and discussed in [21, 45, 27]. The games are often referred to as observational and interventional respectively. Strictly speaking, the interventional game is based on the *do*-operator [36], and only under certain conditions do the marginal and interventional games coincide [49]. For other examples of appropriate games used in ML setting, see the works of [29, 48, 9, 33, 46].

A game value $(N, v) \mapsto h[N, v] \in \mathbb{R}^n$ is a quantification of feature contributions to the model's output when $v \in \{v^{CE}, v^{ME}\}$. Intuitive explanations have been proposed in [45, 21, 8] on how to interpret the game values based on each game. Roughly speaking, conditional game values explain predictions $f(X)$ viewed as a random variable, while marginal game values explain the transformations occurring in the model $f(x)$, sometimes called mechanistic explanations [14]. The work of [8] intuitively describes conditional explanations, also known as observational, as consistent with the data (true-to-the-data) and marginal explanations, also known as interventional, as consistent with the model (true-to-the-model). A popular choice for the game value is the Shapley value [42] (in light of its properties such as symmetry, efficiency, and linearity), but other game values and coalitional values (such as the Owen

[2]In the literature, the conditional and marginal games are typically defined as functions of an observation x instead of X, which makes the corresponding deterministic games.

value [35]) have been investigated in the ML setting before [48, 15, 26]. Some of the articles that describe implementation of Shapley values or their approximates for the above games are [30, 31, 2, 15, 26].

There has been a collection of noted articles devoted to the difference between marginal and conditional Shapley values, a topic which is at the heart of our paper. In [8] the authors argue intuitively that conditional explanations are consistent with the data (true-to-the-data) and marginal explanations are consistent with the model (true-to-the-model). Articles [21, 45] argue that the marginal Shapley value is most appropriate as a mechanistic model explanation [14] (that describes the transformations occurring in the model) because (in light of the null-player property) the marginal game attributes zero value to all predictors that are not explicitly used by the model. The work [27] provides criticisms for both games. The authors indicate that the marginal Shapley value uses out-of-distribution observations, while the conditional one might assign a non-zero attribution to the feature X_i, $i \in N$, which is not used explicitly by the model (but is a proxy of one), and that dropping the player i from the universe of players N leads to different attributions.

In this article, we study explanation maps based on game values defined as linear operators

$$\bar{\mathcal{E}}^{CE}[f;h,X] : f \mapsto h[N, v^{CE}(\cdot;X,f)] \quad \text{and} \quad \bar{\mathcal{E}}^{ME}[f;h,X] : f \mapsto h[N, v^{ME}(\cdot;X,f)] \tag{1.1}$$

on appropriate functional spaces. We investigate the continuity of these operators (on various domains) which helps to illuminate crucial differences between the two games; our theoretical study is motivated by the heuristic concepts of true-to-the-model and true-to-the-data introduced in [8].

Given the above formalism, and motivated by the discussions in [21, 45, 27, 23], we next state several important issues associated with marginal and conditional explanations that we investigate in our work:

(*i*) It is well-known from the Rashomon effect [5] that under predictor dependencies, distinct models that approximate the same data well can have different representations [16]. Consequently, the marginal explanations for models with similar predictions may vary significantly, while conditional explanations will be similar. In other words, the conditional explanations do not depend on the model representation, while the marginal ones do. This property may create practical challenges in applications where models are periodically retrained or when different models are trained on the same data. Moreover, it also has an adverse effect for assessing global feature importance during the modelling process [16].

(*ii*) In light of the curse of dimensionality, computing conditional game values (under predictor dependencies) is often infeasible when the predictor dimension is large, which is the case in many applications; see [19]. There are methods that attempt to approximate conditional games [2, 34] or replace the game with one that mimics the conditioning [29]. It is also common in the literature [30, 29, 9] to design Shapley-based observational attribution methods under assumption of feature independence. Their use, however, often leads to misleading (observational) model explanations [1] that are not consistent with observations.

(*iii*) Given a model with highly dependent features, additive explainers spread any meaningful contributions (of latent variables) across dependent components. This can lead

to rendering their individual explanations extremely minuscule and distort the ranking of individual features based on their explanations; see [27, 1, 23, 32], and §3.3.

In this article, we closely look at issues (i)-(iii) by studying the continuity of suitably defined feature explanation operators. To our knowledge, a rigorous treatment of explanations in a functional analytic setting has never been done before. We believe our work can provide the proper language for understanding when to employ the aforementioned games. Below is a brief summary of technical results presented in our paper.

We set up game-theoretic explainers (1.1) based on a linear game value h in the marginalist form (3.1) as operators. We show that the conditional operator $f \mapsto \bar{\mathcal{E}}^{CE}[f; X, h]$ associated with predictors $X = (X_1, ..., X_n)$ is continuous in $L^2(P_X)$, where P_X is the pushforward measure, and that the Lipschitz bound equals 1 when h is efficient; see Theorem 3.2. Consequently, two models with similar predictions will have similar explanations for the same inputs. This also implies that the conditional explanations for two models with identical predictions are the same regardless of the model representation (that is, they are not impacted by the Rashomon effect [5]); see Lemma 3.4. Similarly, we show that marginal explanations are continuous in a different space $L^2(\tilde{P}_X)$, where $\tilde{P}_X := \frac{1}{2^n}\sum_{S\subseteq N} P_{X_S} \otimes P_{X_{-S}}$; see Theorem 3.8. Thus, any two models that are close in $L^2(\tilde{P}_X)$ (roughly speaking, the models that encode the input-output relationship in a similar way) will have similar marginal explanations.

To study the impact of the Rashomon effect on marginal explanations, we consider the operator $\bar{\mathcal{E}}^{ME}$ on an appropriately-defined space H_X equipped with the $L^2(P_X)$-norm. We show that the marginal game is well-defined on that space if and only if $\tilde{P} \ll P_X$; see Lemma 3.11. For this reason, we focus on the case where $\tilde{P} \ll P_X$. Then, encoding the strength of dependencies by the Radon-Nikodym derivative $r := \frac{d\tilde{P}_X}{dP_X}$, we show that there are two regimes that depend on deviation of $r(x)$ from 1 (measured with respect to P_X). In the first regime, $\bar{\mathcal{E}}^{ME}$ is continuous on H_X but its Lipschitz bound grows as the strength of the dependencies increases. This serves as a precursor of instability because models with very similar predictions may have vastly different explanations. In the second regime, $\bar{\mathcal{E}}^{ME}$ becomes unbounded; see Theorem 3.16, in which case the marginal explanations for two models can be vastly different no matter how close their predictions are.

When $r \in L^\infty(P_X)$, both operators are well-defined on the same space $H_X = L^2(P_X)$. In Proposition 3.22, we establish the approximation bound on the difference between the marginal and conditional maps which highlights the fact that when $r(x)$ deviates from 1 (in the P_X-sense), the marginal explanations will deviate from conditional ones according to this difference; for the general case $r \in L^1(P_X)$ see the approximation bounds in Lemma 3.21.

Finally, for efficient game values, we discuss splits of the explanations energy on dependencies; see Lemma 3.24 that provides the conditions for conservation of energy as well as Example 3.5.

Algorithms such as [31, 30, 15] produce explanation maps that are true-to-the-model as they are implementation-invariant approximators of marginal values. Meanwhile, the path-dependent TreeSHAP method [29] relies on the implementation and fails to be a well-defined map on any of the above spaces [15]. The mitigation strategies for instabilities of marginal explanations based on grouping predictors by dependencies have been discussed in [1, 32].

Structure of the paper. In §2, we introduce the requisite concepts such as the conditional

and marginal games, the notion of game value, and in particular, the Shapley value. The conditional and marginal game operators are set up in §3. Theorems 3.2, 3.8, and 3.16 address the stability of conditional and marginal explanations and highlight their differences. The paper finishes with an appendix containing auxiliary definitions and lemmas.

2. Preliminaries.

2.1. Notation and hypotheses. Throughout this article, we consider the joint distribution (X, Y), where $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ are the predictors, and Y is a response variable with values in $\mathbb{R}$ (not necessarily a continuous random variable). Let the trained model, which estimates the true regressor $\mathbb{E}[Y|X = x]$, be denoted by $f(x)$. We assume that all random variables are defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ a σ-algebra of sets, and $\mathbb{P}$ a probability measure. We let P_X be a pushforward measure of X on $\mathbb{R}^n$ and its support be denoted by $\mathcal{X} := \operatorname{supp}(P_X)$. Similarly, we denote $\mathcal{X}_i := \operatorname{supp}(P_{X_i})$, $i \in \{1, 2, \ldots, n\}$.

Let $S \subseteq N$. Let X_S denote the set of features X_i with $i \in S$ and let $\mathcal{X}_S$ denote its support, where we ignore the predictors' ordering to improve readability. We say that the predictors $X_S = \{X_i\}_{i \in S}$ are independent if $P_{X_S} = \prod_{i \in S} \otimes P_{X_i}$.

Given $\epsilon > 0$, the (L^p, ϵ)-Rashomon set of models about f_* is defined to be the ball of radius ϵ around a given model f_* in the space $L^p(P_X)$, that is, $\{f \in L^p(P_X) : \mathbb{E}[|f_*(X) - f(X)|^p] \leq \epsilon^p\}$. This is a modified version of the definition in [16] which also incorporates the distance from the response variable Y to $f_*(X)$. Finally, the collection of Borel functions on $\mathbb{R}^n$ is denoted by $\mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$.

2.2. Explainability and game theory. The objective of a local model explainer $E(x; f, X) = (E_1, \ldots, E_n)$ is to quantify the contribution of each predictor X_i, $i \in \{1, \ldots, n\}$, to the value of a predictive model $f \in \mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$ at a data instance $x \sim P_X$.

Many promising interpretability techniques utilize ideas from cooperative game theory for constructing explainers. A cooperative game with n players is a set function v that acts on a set of size n, say $N = \{1, 2, \ldots, n\}$, and satisfies $v(\varnothing) = 0$. A game value is a map $v \mapsto h[N, v] \in \mathbb{R}^n$ that determines the worth of each player. See §3.4 for more details.

In the ML setting, the features $X = (X_1, X_2, \ldots, X_n)$ are viewed as n players in an appropriately designed game $S \mapsto v(S; x, X, f)$ associated with the observation $x \sim P_X$, random features X, and model f. The game value $h[N, v]$ then assigns the contributions of each respective feature to the total payoff $v(N; x, X, f)$ of the game at the data instance x.

Two of the most notable games in the ML literature are given by

$$\begin{aligned} &v_*^{CE}(S; x, X, f) = \mathbb{E}[f(X)|X_S = x_S], \quad v_*^{ME}(S; x, X, f) = \mathbb{E}[f(x_S, X_{-S})] \\ &\text{where} \quad v_*^{CE}(\varnothing; x, X, f) = v_*^{ME}(\varnothing; x, X, f) := \mathbb{E}[f(X)] \end{aligned} \tag{2.1}$$

introduced in [43, 30] in the context of the Shapley value [42]

$$\varphi_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n - s - 1)!}{n!} [v(S \cup \{i\}) - v(S)], \quad s = |S|,\ n = |N|. \tag{2.2}$$

The value φ satisfies the axioms of symmetry, linearity and the aforementioned efficiency property (see (SP), (LP) and (EP) in Appendix A). The efficiency property, most appealing to

the ML community, allows for a disaggregation of the payoff $v(N)$ into n parts that represent a contribution to the game by each player: $\sum_{i=1}^n \varphi_i[N, v] = v(N)$.

The games defined in (2.3) are not cooperative since they do not satisfy the condition $v(\varnothing) = 0$. In such a case, the efficiency property reads as $\sum_{i=1}^n \varphi_i[N, v] = v(N) - v(\varnothing)$. See §3.4 for a careful treatment of game values for non-cooperative games.

In this paper, to study game-theoretical explainers in their entirety, we consider random conditional and marginal games given by

$$v^{CE}(S; X, f) = \mathbb{E}[f(X)|X_S], \quad v^{ME}(S; X, f) = \mathbb{E}[f(x_S, X_{-S})]\big|_{x_s = X_S} \tag{2.3}$$

and are related to the deterministic ones in (2.1) via $v^{CE} = v_*^{CE}|_{x=X}$ and $v^{ME} = v_*^{ME}|_{x=X}$. For these games, the corresponding Shapley values $\varphi[N, v^{CE}]$ and $\varphi[N, v^{ME}]$, respectively, are random vectors in $\mathbb{R}^n$. These are well-defined (as operators) when the model f belongs to functional spaces introduced in Section 3.

The deterministic and random Shapley explainers are trivially related as follows:

$$\varphi[N, v^{CE}] = \varphi[N, v_*^{CE}](x)\big|_{x=X}, \quad \varphi[N, v^{ME}] = \varphi[N, v_*^{ME}](x)\big|_{x=X}.$$

This motivates the following definition of a generic random explainer.

Definition 2.1. *Let $X = (X_1, \ldots, X_n)$ be predictors. Suppose $E(\cdot; \cdot, X)$ is a model explainer defined for every $f \in \mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$ and $x \in \mathcal{X}$. Suppose the map $x \mapsto E(x; f, X) \in \mathbb{R}^n$ is Borel. The random model explainer induced by E is defined by $\bar{\mathcal{E}}[f; E, X] := E(X; f, X)$, $f \in \mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$.*

Notice that the map $x \mapsto E(x; f, X)$ in the definition above takes values in $\mathbb{R}^n$ but $\bar{\mathcal{E}}[f; E, X]$ is a random vector of dimension n.

Definition 2.2 (μ-consistency). *Let $X = (X_1, \ldots, X_n), E(\cdot; \cdot, X)$ and $\bar{\mathcal{E}}[\cdot; E, X]$ be as in Definition 2.1. Suppose that $\bar{\mathcal{E}}[f; E, X] \in L^2(\Omega, \mathcal{F}, \mathbb{P})^n$ for every $f \in L^2(\mathbb{R}^n, \mu)$ where μ is a Borel probability measure on $\mathbb{R}^n$. We say that E (and $\bar{\mathcal{E}}$) is μ-consistent if $f \mapsto \bar{\mathcal{E}}[f; E, X]$ is locally Lipshitz continuous, that is, for every $f_* \in L^2(\mu)$ there exists a constant $c = c_{f_*} \geq 0$ such that*

$$\|\bar{\mathcal{E}}[f] - \bar{\mathcal{E}}[f_*]\|_{L^2(\mathbb{P})} \leq c_{f_*} \|f - f_*\|_{L^2(\mu)}, \quad \forall f \in L^2(\mu).$$

The consistency condition guarantees that models that are similar in $L^2(\mu)$, in the sense they are close in $L^2(\mu)$, have similar explanations (up to a scaling constant determined by the bound). For instance, suppose $\mu = P_X$ and $c_{f_*} = 1$. Then, if $\|f_* - f\|_{L^2(P_X)} \leq \epsilon$, that is, the predictions of f and f_* are close to one another within ϵ, then their explanations are also close to each other within ϵ. We further note that if $\bar{\mathcal{E}}$ is linear, then μ-consistency is equivalent to the global Lipshitz continuity with $c(f_*) = \|\bar{\mathcal{E}}\|$ for each $f_* \in L^2(\mu)$.

Remark 2.3. In principle, one can replace the L^2 spaces in Definition 2.2 with the spaces $L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $L^p(\mathbb{R}^n, \mu)$, respectively.

In what follows, when the context is clear, we suppress the explicit dependence of $v \in \{v^{CE}, v^{ME}\}$ on X and f. Furthermore, we will refer to values $\varphi_i[N, v^{ME}]$ and $\varphi_i[N, v^{CE}]$ as marginal and conditional Shapley values.

3. Conditional and marginal game operators. In our work, the game v^{CE} is referred to as conditional and v^{ME} as marginal; see (2.3) for definitions. If predictors X are independent, the two games coincide. In the presence of dependencies, however, the games are very different. The conditional game explores the data by taking into account dependencies, while the marginal game explores the model f in the space of its inputs, ignoring the dependencies. Strictly speaking, the conditional game is determined by the probability measure P_X, while the marginal game is determined by the product probability measures $P_{X_S} \otimes P_{X_{-S}}$, $S \subseteq N$.

The explanations based on these two games have been addressed in the works [45, 21, 8, 33]. These works illustrate that, for certain types of models, the conditional Shapley explanations are consistent with observations while the marginal ones are consistent with the model.

Building upon the aforementioned works, we offer our viewpoint by introducing operators based on the two games whose outputs are explanations viewed as random variables. This construction allows us to better understand the relationships between explanations, the data, and the model.

An appealing property of the marginal and conditional games is that of linearity with respect to models. Specifically, given random features $X = (X_1, X_2, \ldots, X_n)$ and two continuous models f, g we have

$$v(S; X, \alpha \cdot f + g) = \alpha \cdot v(S; X, f) + v(S; X, g), \quad v \in \{v^{CE}, v^{ME}\}.$$

If the game value $h[N, v]$ is also linear, the linearity extends to explanations

$$h[N, v(S; X, \alpha \cdot f + g)] = \alpha \cdot h[N, v(S; X, f)] + h[N, v(S; X, g)], \quad v \in \{v^{CE}, v^{ME}\}$$

on the space of continuous models. To extend the marginal and conditional games to a more general class of models, we consider equivalence classes of models $L^2(\mu)$ for an appropriate Borel probability measure μ, on which the games are well-defined maps. Once the spaces are defined, the linearity of explanations provides a natural approach to obtaining explanations of certain ML ensembles (such as sums of trees) because the construction of explanations focuses on each single term of the ensemble, simplifying the process of determining the appropriate game for a given case.

3.1. Stability theory of single feature explainers based on linear game values. We begin the discussion by introducing linear operators associated with the conditional game and then investigating their properties. A necessary ingredient for constructing such an operator is a linear game value which allows quantifying the contribution of each feature. For simplicity, in this section, we work with the linear game value h in the (marginalist) form

$$h_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(S, n)\big[v(S \cup \{i\}) - v(S)\big], \quad i \in N = \{1, 2, \ldots, n\}. \tag{3.1}$$

Such game values are determined by weights $w(S, n)$ where S is a proper subset of N. Notice that the Shapley value (2.2) is of the form above. Indeed, game values of this form satisfy desirable properties such as linearity (LP) and the null-player property (NPP); see A.

Definition 3.1. *Let $X = (X_1, \ldots, X_n)$ be defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and $h[N, v]$ be given by (3.1).*

(i) The conditional game operator $\mathcal{E}^{CE} : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to L^2(\Omega, \mathcal{F}, \mathbb{P})^n$ *associated with* h, X *is defined by*

$$\mathcal{E}_i^{CE}[Z; h, X] := \sum_{S \subseteq N \setminus \{i\}} w(S, n)\big[\mathbb{E}[Z|X_{S\cup\{i\}}] - \mathbb{E}[Z|X_S]\big], \quad i \in N, \tag{3.2}$$

where we set $\mathbb{E}[Z|X_\varnothing] := \mathbb{E}[Z]$.

(ii) The pullback conditional game operator $\bar{\mathcal{E}}^{CE} : L^2(P_X) \to L^2(\Omega, \mathcal{F}, \mathbb{P})^n$ *associated with* h, X *is defined by*

$$\bar{\mathcal{E}}^{CE}[f; h, X] := h[N, v^{CE}(\cdot; X, f)].$$

Throughout this section, for the ease of notation, we denote the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ by $L^2(\mathbb{P})$ and assume that $X = (X_1, \dots, X_n)$ is a random vector defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

Theorem 3.2 (properties). *Let* h, X *and* $\mathcal{E}^{CE}$ *be as in Definition 3.1.*

(i) $\mathcal{E}_i^{CE}$ *is a bounded linear, self-adjoint operator satisfying*

$$\|\mathcal{E}_i^{CE}[Z; h, X]\|_{L^2(\mathbb{P})} \le \Big(\sum_{S \subseteq N\setminus\{i\}} |w(S, n)|\Big)\|Z\|_{L^2(\mathbb{P})}. \tag{3.3}$$

(ii) Let $X_i \in L^2(\mathbb{P})$. *If* $X_i \perp\!\!\!\perp X_{N\setminus\{i\}}$ *and (NN) holds, then* $\|\mathcal{E}_i^{CE}\| = \sum_{S\subseteq N\setminus\{i\}} |w(S, n)|$.

(iii) $\{Z \in L^2(\mathbb{P}) : Z \perp\!\!\!\perp X\} \subseteq \{Z \in L^2(\mathbb{P}) : \mathbb{E}[Z|X_{S\cup\{i\}}] = \mathbb{E}[Z|X_S],\ S \subseteq N \setminus \{i\}\} \subseteq \mathrm{Ker}(\mathcal{E}_i^{CE})$, $i \in N$.

(iv) $\{Z \in L^2(\mathbb{P}) : Z \perp\!\!\!\perp X\} \subseteq \{Z \in L^2(\mathbb{P}) : \mathbb{E}[Z|X] = const\ \mathbb{P}\text{-a.s.}\} \subseteq \mathrm{Ker}(\mathcal{E}^{CE})$.

(v) $\mathrm{Ker}(\mathcal{E}^{CE}) = \{Z \in L^2(\mathbb{P}) : \mathbb{E}[Z|X] = const\ \mathbb{P}\text{-a.s.}\}$ *if* h *satisfies axiom (TPG).*

(vi) If h *satisfies the efficiency property (EP), then* $\sum_{i=1}^n \mathcal{E}_i^{CE}(Z) = \mathbb{E}[Z|X] - \mathbb{E}[Z]$.

Proof. The linearity of the operator is a consequence of the linearity of the expected value. To estimate the norm, observe that

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[Z|X_{S\cup\{i\}}] - \mathbb{E}[Z|X_S]\big]^2 &= \mathbb{E}\big[\mathbb{E}[Z|X_{S\cup\{i\}}] - \mathbb{E}[\mathbb{E}[Z|X_{S\cup\{i\}}]|X_S]\big]^2 \\ &\le Var(\mathbb{E}[Z|X_{S\cup\{i\}}]) \le \|Z\|^2_{L^2(\mathbb{P})}.\end{aligned}$$

Hence $\|\mathcal{E}_i^{CE}[Z]\|_{L^2(\mathbb{P})} \le \sum_{S\subseteq N: i\in N} |w(S, n)| \cdot \|Z\|_{L^2(\mathbb{P})}$ which gives the estimate (3.3).

Next, note that the operator $\mathcal{E}_i^{CE}$ can be expressed as

$$\mathcal{E}_i^{CE} = \sum_{S\subseteq N\setminus\{i\}} w(S, n)\big(P_{S\cup\{i\}} - P_S\big), \tag{3.4}$$

where P_S is the orthogonal projection operator with values in $L^2(\Omega, \sigma(X_S), \mathbb{P})$ defined by $P_S[Z] := \mathbb{E}[Z|X_S]$. Since P_S and $I - P_S$ project on orthogonal spaces, we have

$$\langle P_S[Z_1], Z_2\rangle = \langle P_S[Z_1], P_S[Z_2]\rangle = \langle Z_1, P_S[Z_2]\rangle \quad \text{for all} \quad Z_1, Z_2 \in L^2(\Omega, \mathcal{F}, \mathbb{P}),$$

and hence, using (3.4), we conclude that $\mathcal{E}_i^{CE} = \mathcal{E}_i^{CE*}$. This proves (i).

Suppose that $w(S,n) \geq 0$ for all $S \subseteq N$, and X_i is independent of $X_{N\setminus\{i\}}$. Then

$$\|\mathcal{E}_i^{CE}[X_i - \mathbb{E}[X_i]]\|_{L^2(\mathbb{P})} = \Big(\sum_{S\subseteq N\setminus\{i\}} w(S,n)\Big)\|X_i - \mathbb{E}[X_i]\|_{L^2(\mathbb{P})} \geq \|\mathcal{E}_i^{CE}\| \cdot \|X_i - \mathbb{E}[X_i]\|_{L^2(\mathbb{P})}$$

which implies (ii).

The first inclusion in (iii) is obvious and the second one follows from (3.2). Part (iv) follows from (iii) because $\mathrm{Ker}(\mathcal{E}^{CE}) = \bigcap_{i=1}^n \mathrm{Ker}(\mathcal{E}_i^{CE})$ contains subspaces

$$\bigcap_{i=1}^n \{Z \in L^2(\mathbb{P}) : \mathbb{E}[Z|X_{S\cup\{i\}}] = \mathbb{E}[Z|X_S],\ S \subseteq N\setminus\{i\}\} \supseteq$$
$$\{Z \in L^2(\mathbb{P}) : \mathbb{E}[Z|X] = const\ \mathbb{P}\text{-a.s.}\} \supseteq \{Z \in L^2(\mathbb{P}) :\ Z \perp\!\!\!\perp X\}.$$

Suppose next the (TPG) property, i.e. (A.5), holds. Then for any $Z \in \mathrm{Ker}(\mathcal{E}^{CE})$:

$$0 = \sum_{i=1}^n |\mathcal{E}_i^{CE}[Z - \mathbb{E}[Z]]| \geq g(|\mathbb{E}[Z|X] - \mathbb{E}[Z]|, |N|) \geq 0.$$

■

Since $g(a,n) = 0$ if and only if $a = 0$, we obtain $\mathbb{E}[Z|X] = \mathbb{E}[Z]$ $\mathbb{P}$-a.s. This concludes the proof of (v). The property (vi) follows directly from the efficiency property (EP).

Remark 3.3. An immediate consequence of Theorem 3.2(i)-(iv) is the following stronger inequality $\|\mathcal{E}_i^{CE}[Z;h,X]\|_{L^2(\mathbb{P})} \leq \big(\sum_{S\subseteq N\setminus\{i\}} |w(S,n)|\big)\|Z - \mathbb{E}[Z]\|_{L^2(\mathbb{P})}$.

We next present two corollaries to Theorem 3.2. The first one states that the conditional explanations of the regressor and the response variable coincide.

Corollary 3.4. *Let $Y \in L^2(\Omega,\mathcal{F},\mathbb{P})$. Set $\epsilon := Y - \mathbb{E}[Y|X]$. Then*

$$\mathcal{E}^{CE}[Y;h,X] = \mathcal{E}^{CE}[\mathbb{E}[Y|X];h,X], \quad \mathcal{E}^{CE}[\epsilon;h,X] = 0.$$

Proof. Follows immediately from that fact that $\epsilon \in \mathrm{Ker}(\mathcal{E}^{CE})$ due to Theorem 3.2(iv). ■

Corollary 3.5. *Let h, X, $\bar{\mathcal{E}}^{CE}$ be as in Definition 3.1.*

(i) *The operator $\bar{\mathcal{E}}^{CE}$ is a bounded linear operator satisfying*

$$\|\bar{\mathcal{E}}^{CE}[f_1;h,X] - \bar{\mathcal{E}}^{CE}[f_2;h,X]\|_{L^2(\mathbb{P})^n} \leq C\|f_1 - f_2\|_{L^2(P_X)}.$$

Here $C := \sqrt{n}\max_i(C_i)$ where C_i is the constant on the right-hand side of (3.3).

(ii) *For a game value h of the form (3.1) which satisfies (NN) and the efficiency property (EP), the Lipschitz inequality from (i) can be improved as*

$$\|\bar{\mathcal{E}}^{CE}[f_1;h,X] - \bar{\mathcal{E}}^{CE}[f_2;h,X]\|_{L^2(\mathbb{P})^n} \leq \|f_1 - f_2\|_{L^2(P_X)}. \tag{3.5}$$

(iii) *One has $\mathrm{Ker}(\bar{\mathcal{E}}^{CE}) \supseteq \{f \in L^2(P_X) : f = const\ P_X\text{-a.s.}\}$ with equality achieved if h satisfies (TPG).*

Proof. Parts (i) and (iii) of the corollary follow immediately from $\bar{\mathcal{E}}^{CE}[f] = \mathcal{E}^{CE}[f(X)]$, and parts (i), (iv) and (v) of Theorem 3.2. Part (ii) is more subtle: By Theorem 3.2(vi), the efficiency property puts a constraint on $\bar{\mathcal{E}}^{CE}[f] = \big(\bar{\mathcal{E}}_1^{CE}[f], \ldots, \bar{\mathcal{E}}_n^{CE}[f]\big)$; its components should add up to $f(X) - \mathbb{E}[f(X)]$. As we shall see, this constraint allows for a better estimation of the norm of this vector. There is no loss of generality in assuming that $\mathbb{E}[f(X)] = 0$ since constant functions lie in the kernel. Now it suffices to establish $\|\bar{\mathcal{E}}^{CE}[f]\|_{L^2(\mathbb{P})^n} \leq \|f\|_{L^2(P_X)}$. Notice that

$$
\begin{aligned}
\|f\|^2_{L^2(P_X)} = \|f(X)\|^2_{L^2(\mathbb{P})} = \langle f(X), \sum_{i=1}^n \bar{\mathcal{E}}_i^{CE}[f]\rangle_{L^2(\mathbb{P})} &= \sum_{i=1}^n \langle f(X), \bar{\mathcal{E}}_i^{CE}[f]\rangle_{L^2(\mathbb{P})} \\
&= \sum_{i=1}^n \Big(\sum_{S\subseteq N\setminus\{i\}} w(S,n) \langle f(X), \mathbb{E}[f(X)|X_{S\cup\{i\}}] - \mathbb{E}[f(X)|X_S]\rangle_{L^2(\mathbb{P})} \Big) \\
&= \sum_{i=1}^n \Big(\sum_{S\subseteq N\setminus\{i\}} w(S,n) \|\mathbb{E}[f(X)|X_{S\cup\{i\}}] - \mathbb{E}[f(X)|X_S]\|^2_{L^2(\mathbb{P})} \Big).
\end{aligned} \tag{3.6}
$$

The last equality is based on interpreting conditional expectation as orthogonal projections which indicates that the inner products $\langle f(X) - \mathbb{E}[f(X)|X_S], \mathbb{E}[f(X)|X_S]\rangle_{L^2(\mathbb{P})}$, $\langle f(X) - \mathbb{E}[f(X)|X_{S\cup\{i\}}], \mathbb{E}[f(X)|X_{S\cup\{i\}}]\rangle_{L^2(\mathbb{P})}$ and $\langle \mathbb{E}[f(X)|X_{S\cup\{i\}}] - \mathbb{E}[f(X)|X_S], \mathbb{E}[f(X)|X_S]\rangle_{L^2(\mathbb{P})}$ are all zero. The number $\|f\|^2_{L^2(P_X)}$, as described above, is not smaller than $\|\bar{\mathcal{E}}^{CE}[f]\|^2_{L^2(\mathbb{P})^n}$ due to:

$$
\begin{aligned}
\|\bar{\mathcal{E}}^{CE}[f]\|^2_{L^2(\mathbb{P})^n} &= \sum_{i=1}^n \|\bar{\mathcal{E}}_i^{CE}[f]\|^2_{L^2(\mathbb{P})} \\
&= \sum_{i=1}^n \| \sum_{S\subseteq N\setminus\{i\}} w(S,n) \big[\mathbb{E}[f(X)|X_{S\cup\{i\}}] - \mathbb{E}[f(X)|X_S]\big] \|^2_{L^2(\mathbb{P})} \\
&\leq \sum_{i=1}^n \Big(\Big(\sum_{S\subseteq N\setminus\{i\}} w(S,n) \Big) \Big(\sum_{S\subseteq N\setminus\{i\}} w(S,n) \|\mathbb{E}[f(X)|X_{S\cup\{i\}}] - \mathbb{E}[f(X)|X_S]\|^2_{L^2(\mathbb{P})} \Big) \Big) \\
&= \sum_{i=1}^n \Big(\sum_{S\subseteq N\setminus\{i\}} w(S,n) \|\mathbb{E}[f(X)|X_{S\cup\{i\}}] - \mathbb{E}[f(X)|X_S]\|^2_{L^2(\mathbb{P})} \Big);
\end{aligned} \tag{3.7}
$$

where on the second line we used Cauchy-Schwarz along with $w(S,n) \geq 0$ while the third line relies on $\sum_{S\subseteq N\setminus\{i\}} w(S,n) = 1$ which follows from the efficiency property. ∎

Remark 3.6. Arguments in the proof of Corollary 3.5(ii) show that if conditions $w(S,n) \geq 0$ and the efficiency are satisfied, the conditional operator (in light of Corollary 3.4) also satisfies the sharper bound $\|\mathcal{E}^{CE}[Z_1 - Z_2; h, X] - \mathcal{E}^{CE}[Z_2; h, X]\|_{L^2(\mathbb{P})^n} \leq \|Z_1 - Z_2\|_{L^2(\mathbb{P})}$.

Corollary 3.5 implies that for two distinct models $f_1(x)$, $f_2(x)$ that approximate the data well, the conditional explanations are consistent with those of the data.

We next take a similar approach in constructing an operator based on the marginal game.

To choose an appropriate space of models, note that for any bounded $f \in \mathcal{B}(\mathbb{R}^n)$ we have

$$\mathbb{E}\big[v^{ME}(S;X;f)\big] = \int f(x_S, x_{-S})[P_{X_S} \otimes P_{X_{-S}}](dx_S, dx_{-S}).$$

Since the marginal explanations based on the game value (3.1) are linear combinations of $v^{ME}(S;X;f)$, $S \subseteq N$, natural domains for the marginal operator are the spaces $L^q(\tilde{P}_X)$, $q \geq 1$, with the corresponding co-domains being $L^q(\mathbb{P})$, where

$$\tilde{P}_X := \frac{1}{2^n} \sum_{S \subseteq N} P_{X_S} \otimes P_{X_{-S}} \tag{3.8}$$

with the corresponding L^q-norm

$$\|f\|^q_{L^q(\tilde{P}_X)} := \frac{1}{2^n} \sum_{S \subseteq N} \int f^q(x_S, x_{-S})[P_{X_S} \otimes P_{X_{-S}}](dx_S, dx_{-S}),$$

where we ignore the variable ordering in f to ease the notation, and we assign $P_{X_\varnothing} \otimes P_X = P_X \otimes P_{X_\varnothing} = P_X$. In what follows, we develop the L^2-theory for the marginal explanations.

Definition 3.7. *Let h, X be as in Definition 3.1. The marginal game operator $\bar{\mathcal{E}}^{ME} : L^2(\tilde{P}_X) \to L^2(\Omega, \mathcal{F}, \mathbb{P})^n$ associated with h, X is defined by*

$$\bar{\mathcal{E}}^{ME}[f;h,X] := h[N, v^{ME}(\cdot\,;X,f)]. \tag{3.9}$$

Theorem 3.8 (properties). *Let X, h, and $(\bar{\mathcal{E}}^{ME}, L^2(\tilde{P}_X))$ be as in Definition 3.7.*
(i) $\bar{\mathcal{E}}^{ME}_i$ is a well-defined, bounded linear operator satisfying

$$\|\bar{\mathcal{E}}^{ME}_i[f;h,X]\|_{L^2(\mathbb{P})} \leq 2^{\frac{n+1}{2}} \Big(\sum_{S \subseteq N \setminus \{i\}} w^2(S,n) \Big)^{\frac{1}{2}} \|f\|_{L^2(\tilde{P}_X)}.$$

(ii) $\{f \in L^2(\tilde{P}_X) : f = const\ \tilde{P}_X\text{-a.s.}\} \subseteq \mathrm{Ker}(\bar{\mathcal{E}}^{ME})$.
(iii) If axiom (TPG) holds, then $\mathrm{Ker}(\bar{\mathcal{E}}^{ME}) \subseteq \{f \in L^2(\tilde{P}_X) : f = const\ P_X\text{-a.s.}\}$.
(iv) If axiom (TPG) holds and $\tilde{P}_X \ll P_X$, $\mathrm{Ker}(\bar{\mathcal{E}}^{ME}) = \{f \in L^2(\tilde{P}_X) : f = const\ \tilde{P}_X\text{-a.s.}\}$.
(v) If $f(x) = f(x_{N\setminus\{i\}})$ for some $i \in N$, then i is a null player for $v^{ME}(\cdot; X, f)$.
(vi) $\{f \in L^2(\tilde{P}_X) : f(x) = f(x_{N\setminus\{i\}})\ \tilde{P}_X\text{-a.s.}\} \subseteq \mathrm{Ker}(\bar{\mathcal{E}}^{ME}_i)$.
(vii) If h satisfies the efficiency property (EP), then $\sum_{i=1}^n \bar{\mathcal{E}}^{ME}_i[f] = f(X) - \mathbb{E}[f(X)]$.

Proof. If $f = f_*$ $\tilde{P}_X$-a.s., then for any $S \subseteq N$

$$v^{ME}(S \cup \{i\}; X, f) = v^{ME}(S \cup \{i\}; X, f_*)\ \mathbb{P}\text{-a.s.}$$

which implies, in view of (3.1), that $\bar{\mathcal{E}}^{ME}$ is well-defined on $L^2(\tilde{P}_X)$.

Now let $\bar{f}_S(x_S) = \mathbb{E}[f(x_S, X_{-S})]$. Then

$$\begin{aligned}
\|\bar{\mathcal{E}}_i^{ME}[f]\|_{L^2(\mathbb{P})} &\leq \sum_{S\subseteq N\setminus\{i\}} |w(S,n)| \cdot \|\bar{f}_{S\cup\{i\}}(X_{S\cup\{i\}}) - \bar{f}_S(X_S)\|_{L^2(\mathbb{P})} \\
&\leq \Big(\sum_{S\subseteq N\setminus\{i\}} w^2(S,n)\Big)^{\frac{1}{2}} \Big(\sum_{S\subseteq N\setminus\{i\}} \|\bar{f}_{S\cup\{i\}}(X_{S\cup\{i\}}) - \bar{f}_S(X_S)\|^2_{L^2(\mathbb{P})}\Big)^{\frac{1}{2}} \\
&= \Big(\sum_{S\subseteq N\setminus\{i\}} w^2(S,n)\Big)^{\frac{1}{2}} \Big(2\sum_{S\subseteq N} \|f\|^2_{L^2(P_{X_S}\otimes P_{X_{-S}})}\Big)^{\frac{1}{2}} \\
&= 2^{\frac{n+1}{2}} \Big(\sum_{S\subseteq N\setminus\{i\}} w^2(S,n)\Big)^{\frac{1}{2}} \cdot \|f\|_{L^2(\tilde{P}_X)}.
\end{aligned}$$

which establishes (i).

We next prove (ii). Suppose $f = c$ $\tilde{P}_X$-a.s. for some constant $c \in \mathbb{R}$. Let $f_*(x) := c$ for each $x \in \mathbb{R}^n$. Note that for any $S \subseteq N$, including $S = \varnothing$, we have

$$v^{ME}(S\cup\{i\}; X, f_*) - v^{ME}(S; X, f_*) = 0 \quad \mathbb{P}\text{-a.s.},$$

and from (3.1) it follows that $\bar{\mathcal{E}}^{ME}[f_*] = 0 \in L^2(\mathbb{P})$. Note that $f = f_*$ $\tilde{P}_X$-a.s. and hence, using the fact that $\bar{\mathcal{E}}^{ME}$ is well-defined, we conclude that $\bar{\mathcal{E}}^{ME}[f] = 0 \in L^2(\mathbb{P})$ which establishes (ii).

Suppose that $f \in \mathrm{Ker}(\bar{\mathcal{E}}^{ME})$ and (A.5) holds. Then

$$0 = \sum_{i=1}^n |\bar{\mathcal{E}}_i^{ME}[f - \mathbb{E}[f(X)]| \geq g(|f(X) - \mathbb{E}[f(X)]|, n) \geq 0,$$

and hence $f = \mathbb{E}[f(X)]$ P_X-a.s., which proves (iii).

Suppose that $\tilde{P}_X \ll P_X$ and (A.5) holds. Then for any constant $c \in \mathbb{R}$, $f = c$ P_X-a.s. implies $f = c$ $\tilde{P}_X$-a.s. and hence, using (ii) and (iii), we obtain (iv).

Next, if $f(x) = f(x_{N\setminus\{i\}})$, then $\bar{f}_{S\cup\{i\}}(X_{S\cup\{i\}}) = \bar{f}_S(X_S)$ and hence $\bar{\mathcal{E}}_i^{ME}[f] = 0 \in L^2(\mathbb{P})$, which gives (v) and (vi). Property (vii) follows directly from the efficiency property (EP). ■

Remark 3.9. A consequence of Theorem 3.8(i)-(ii) is the following stronger inequality

$$\|\bar{\mathcal{E}}_i^{ME}[f; h, X]\|_{L^2(\mathbb{P})} \leq 2^{\frac{n+1}{2}} \Big(\sum_{S\subseteq N\setminus\{i\}} w^2(S,n)\Big)^{\frac{1}{2}} \|f - \tilde{f}_0\|_{L^2(\tilde{P}_X)}, \quad \tilde{f}_0 := \mathbb{E}_{x\sim\tilde{P}_X}[f(x)].$$

Theorem 3.8(i) states that the marginal operator is bounded in $L^2(\tilde{P}_X)$ and hence the marginal explanations are continuous in $L^2(\tilde{P}_X)$. Under dependencies in predictors, however, two models that are close in $L^2(P_X)$ may yield (as we will see) marginal explanations that are far apart in $L^2(\tilde{P}_X)$, which may cause the map $(X, f) \mapsto \bar{\mathcal{E}}^{ME}[f; X]$ to be unbounded on some other domains; see the discussion below in §3.2.

Remark 3.10. The theory we developed in §3.1 views explanations as maps from a space of models to a space of random variables. While the intuitive notions of true-to-the-model and true-to-the-data introduced in [8] are not equivalent to the continuity in $L^2(\tilde{P}_X)$ and $L^2(P_X)$,

respectively, they are related. Roughly speaking, for explanations to be true-to-the-data, it is necessary for the explanation map to be continuous in $L^2(P_X)$, and to be true-to-the-model continuity in $L^2(\tilde{P}_X)$ is required. Below we present a simple example illustrating that marginal explanations depend on the model representation, while the conditional ones do not.

Example 3.1. Let $X = (X_1, X_2, X_3)$ with $\mathbb{E}[X_i] = 0$. Suppose that $X_i = Z + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \delta)$, $i \in \{1, 2\}$, for some small $\delta > 0$, where $Z \sim \mathcal{N}(0, 1)$. Also suppose that $\epsilon_1, \epsilon_2, Z, X_3$ are independent, and let the response variable be $Y = f_0(X) := X_1 + X_2 + X_3$.

Note that there are many good models defined on $\mathcal{X} = \widetilde{\mathcal{X}} = \mathbb{R}^3$ that represent the same data in $L^2(P_X)$-sense. For instance, consider

$$f_\alpha(x) = (1+\alpha)x_1 + (1-\alpha)x_2 + x_3, \quad x \in \mathbb{R}^3,\ \alpha \in [0, 1],$$

in which case the response variable can be expressed by $Y = f_\alpha(X) + \epsilon_\alpha$ where $\epsilon_\alpha := \alpha(\epsilon_2 - \epsilon_1)$ with $\|\epsilon_\alpha\|_{L^2(\mathbb{P})} \leq \sqrt{2}\delta$. Note that for any $\alpha \in [0, 1]$ the model f_α satisfies:

$$f_\alpha \in L^2(\widetilde{P}_X), \quad \|f_\alpha - f_0\|_{L^2(P_X)} = \sqrt{2}\delta\alpha, \quad \alpha \leq \|f_\alpha - f_0\|_{L^2(\tilde{P}_X)} < \infty.$$

Consider next the conditional explanations based on Shapley value $h = \varphi$. Direct computations of the explanations for the response variable give:

$$\begin{aligned}
\mathcal{E}_1^{CE}[Y; \varphi, X] &= \bar{\mathcal{E}}_1^{CE}[f_0; \varphi, X] = \frac{1}{2}\big(2X_1 + \mathbb{E}[X_2|X_1] - \mathbb{E}[X_1|X_2]\big) = X_1 + O(\delta),\\
\mathcal{E}_1^{CE}[Y; \varphi, X] &= \bar{\mathcal{E}}_2^{CE}[f_0; \varphi, X] = \frac{1}{2}\big(2X_2 + \mathbb{E}[X_1|X_2] - \mathbb{E}[X_2|X_1]\big) = X_2 + O(\delta),\\
\mathcal{E}_1^{CE}[Y; \varphi, X] &= \bar{\mathcal{E}}_3^{CE}[f_0; \varphi, X] = X_3.
\end{aligned}$$

Since $\epsilon_\alpha \perp\!\!\!\perp X_3$, the conditional Shapley explanations for f_α are given by

$$\bar{\mathcal{E}}_1^{CE}[f_\alpha; \varphi, X] = X_1 + O(\delta), \quad \bar{\mathcal{E}}_2^{CE}[f_\alpha; \varphi, X] = X_2 + O(\delta), \quad \bar{\mathcal{E}}_3^{CE}[f_\alpha; \varphi, X] = X_3. \tag{3.10}$$

Thus, we get the same conditional explanations for all models f_α as $\delta \to 0$.

On the other hand, computing marginal expectations, we obtain

$$\bar{\mathcal{E}}_1^{ME}[f_\alpha; \varphi, X] = (1+\alpha)X_1, \quad \bar{\mathcal{E}}_2^{ME}[f_\alpha; \varphi, X] = (1-\alpha)X_2, \quad \bar{\mathcal{E}}_3^{ME}[f_\alpha; \varphi, X] = X_3. \tag{3.11}$$

Comparing equations (3.10) and (3.11), we see that the conditional Shapley values for predictors X_1, X_2 are independent of the representative model up to small additive noise, while that is not the case for the marginal ones.

3.2. Stability of marginal explanations on a space equipped with $L^2(P_X)$-norm. The objective of this section is to investigate when the marginal explanations behave as the conditional ones. That is, we will determine when we can expect two models that have similar predictions to have similar marginal explanations, and how the dependencies in features impact dissimilarity. To answer these questions, it is necessary to investigate the stability of marginal explanations on a space equipped with $L^2(P_X)$-norm.

If one attempts to equip the space $L^2(\tilde{P}_X)$ with the $L^2(P_X)$-norm, then the marginal game operator may not always be well-defined or bounded; see Theorem 3.13 and Theorem 3.16. To understand this, define the following space:

$$H_X := \left(\left\{ [f] : [f] = \left\{ \tilde{f} : \tilde{f} = f \ P_X\text{-a.s. and } \int |\tilde{f}(x)|^2 \tilde{P}_X(dx) < \infty \right\} \right\}, \|\cdot\|_{L^2(P_X)} \right) \hookrightarrow L^2(P_X). \tag{3.12}$$

Note that either H_X contains exactly the same elements as $L^2(\tilde{P}_X)$ or some elements of $L^2(\tilde{P}_X)$ are placed in the same equivalence class of H_X. Strictly speaking, H_X is a quotient space of $L^2(\tilde{P}_X)$ modulo $H_X^0 := \{f \in L^2(\tilde{P}_X) : \|f\|_{L^2(P_X)} = 0\}$ equipped with the $L^2(P_X)$-norm; keep in mind that, as $P_X \ll \tilde{P}_X$, if $f_1 = f_2$ $\tilde{P}_X$-a.s., then $f_1 = f_2$ P_X-almost surely.

It turns out, as the lemma below states, that the absolute continuity of $\tilde{P}_X$ w.r.t. P_X is a necessary and sufficient condition for the marginal game to be a well-defined map on H_X.

Lemma 3.11. *The map $f \in H_X \mapsto \{v^{ME}(S; X, f)\}_{S \subseteq N} \in (L^2(\mathbb{P}))^{2^n}$ is well-defined if and only if $\tilde{P}_X \ll P_X$. Consequently, $(\bar{\mathcal{E}}^{ME}[\cdot; h, X], H_X)$ is well-defined for every linear game value h if and only if $\tilde{P}_X \ll P_X$.*

Proof. Suppose that the map $[f] \in H_X \mapsto v^{ME}(S; f, X) \in L^2(\mathbb{P})$ is well-defined for every $S \subseteq N$. Suppose that $\tilde{P}_X \not\ll P_X$. Then there exists $A \subset \mathcal{B}(\mathbb{R}^n)$ and $S \subset N$ such that $P_X(A) = 0$ and $P_{X_S} \otimes P_{X_{-S}}(A) > 0$. Set $f_*(x) = 1_A(x)$. Since $\|f_*\|_{H_X}^2 = P_X(A) = 0$, we conclude $f \in [0]_{H_X}$. Hence $v^{ME}(S; f_*, X) = v^{ME}(S; 0, X)$ $\mathbb{P}$-almost surely. This however leads to a contradiction because

$$\mathbb{E}[v^{ME}(S; f_*, X)] = \int \int 1_A(x) \, P_{X_S}(dx_S) P_{X_{-S}} dx_{-S}) = P_{X_S} \otimes P_{X_{-S}}(A) > 0. \qquad \blacksquare$$

Suppose that $\tilde{P}_X \ll P_X$. Any $f \in [0]_{H_X}$ is P_X-almost surely zero; it is thus almost surely zero with respect to $\tilde{P}_X$, in particular with respect to any probability measure $P_{X_S} \otimes P_{X_{-S}}$ where $S \subseteq N$. This implies that $v^{ME}(S; f, X) \in L^2(\mathbb{P})$ is zero:

$$\mathbb{E}\big[|v^{ME}(S; X; f)|\big] = \int |f(x_S, x_{-S})| \, [P_{X_S} \otimes P_{X_{-S}}](dx_S, dx_{-S}) = 0.$$

In other words, the above lemma states that if the density of $\tilde{P}_X$ with respect to P_X exists, then the game value as an operator on H_X is well-defined. A geometric consequence of the above lemma is given in Remark 3.12.

Remark 3.12. In the supplemental material, measures P_X and $\tilde{P}_X$ are compared more carefully, and the relation between the continuity condition $\tilde{P}_X \ll P_X$ and the shape of the support of P_X is investigated. Indeed, the condition holds if the values that predictors assume are in a sense "heterogeneous" while it fails if the predictors live on a "complicated" lower-dimensional submanifold.

Given the lemma above, it is not surprising that the absolute continuity also comes up with regard to the marginal operator.

Theorem 3.13 (well-posedness). *Let h, X be as in Definition 3.1.*

(*i*) *Suppose* $\tilde{P}_X \ll P_X$. *Then* $H_X \cong (L^2(\tilde{P}_X), \|\cdot\|_{L^2(P_X)})$ *and* $(\bar{\mathcal{E}}^{ME}[\cdot; h, X], H_X)$ *acting via the formula* (3.9) *is well-defined.*

(*ii*) *Suppose* $\tilde{P}_X \not\ll P_X$. *Then, for each* $[f] \in H_X$ *there exist* $f_1, f_2 \in [f]$, *such that* $\|f_1 - f_2\|_{L^2(\tilde{P}_X)} \neq 0$. *Consequently,* $H_X \cong (L^2(\tilde{P}_X)/H_X^0, \|\cdot\|_{L^2(P_X)})$, *and* $(\bar{\mathcal{E}}^{ME}[\cdot; h, X], H_X)$ *is well-defined if and only if*

$$H_X^0 = \left\{ f \in L^2(\widetilde{P}_X) : \|f\|_{L^2(P_X)} = 0 \right\} \subseteq \mathrm{Ker}(\bar{\mathcal{E}}^{ME}[\cdot; h, X], L^2(\widetilde{P}_X)).$$

Proof. Given the definition of $\widetilde{P}_X$ in (3.8), one has $P_X(A) \leq 2^n \cdot \widetilde{P}_X(A)$ for any Borel subset of $\mathbb{R}^n$. In particular, $P_X \ll \widetilde{P}_X$ and there exists a well-defined bounded linear map $\tilde{I} : L^2(\widetilde{P}_X) \to L^2(P_X)$ that takes the $L^2(\widetilde{P}_X)$-class of a function to its $L^2(P_X)$-class; notice that $\tilde{I}$ is not necessarily injective or surjective in general. Observe that H_X is the image of $\tilde{I}$; and recall that $\mathrm{Im}(\tilde{I})$ can be identified with $L^2(\widetilde{P}_X)/\mathrm{Ker}(\tilde{I})$ as vector spaces. Thus, the well-defined operator $(\bar{\mathcal{E}}^{ME}, L^2(\widetilde{P}_X))$ can be pushforwarded via $\tilde{I}$ to a well-defined operator $(\bar{\mathcal{E}}^{ME}, H_X)$ if and only if

$$H_X^0 = \mathrm{Ker}(\tilde{I}) = \left\{ f \in L^2(\widetilde{P}_X) : \|f\|_{L^2(P_X)} = 0 \right\} \subseteq \mathrm{Ker}(\bar{\mathcal{E}}^{ME}, L^2(\widetilde{P}_X)).$$

Part (*ii*) describes the situation where H_X^0 is non-trivial while part (*i*) addresses the case where $\tilde{I}$ is an embedding onto the subspace H_X. The latter happens precisely when $\widetilde{P}_X \ll P_X$. ∎

Part (*ii*) of Theorem 3.13 states that, even if $\tilde{P}_X \not\ll P_X$, the marginal operator on H_X may still be well-defined if H_X^0 is in the kernel of the marginal operator on $L^2(\tilde{P}_X)$ since functions in equivalence classes of H_X^0 when plugged into the formula (3.1) yield zero explanations. In such a situation, the linear combination of terms $v^{ME}(S; f, X)$ encoded by h gives rise to a well-defined map on H_X even though at least one assignment $[f] \mapsto v^{ME}(S; f, X)$ should be ill-posed, as according to Lemma 3.11.

Example 3.2. Consider $h = \varphi$. Let $X = (X_1, X_2)$ satisfy $X_2 = g(X_1) + Z$ where Z is a bounded random variable independent of X_1 and g is continuous. Suppose that the supports of X_1, X_2 are $\mathcal{X}_1 = \mathcal{X}_2 = \mathbb{R}$, and that $|Z| \leq M$. In this case, $\mathcal{X} \subseteq \{(x_1, x_2) : x_1 \in \mathbb{R}, |x_2 - g(x_1)| \leq M\}$ where $\mathcal{X}$ is the support of (X_1, X_2), and hence the complement $\mathcal{X}^C$ is a non-empty open set. Pick any open rectangle $R = (a, b) \times (c, d) \subset \mathcal{X}^C$ and set $f_R(x) := \mathbb{1}_R(x) \in L^2(\widetilde{P}_X)$.

Then, using the fact $P_X(R) = 0$, we obtain $\mathbb{P}$-a.s.

$$\varphi_1[v^{ME}(\cdot; X, f_R)] = -\varphi_2[v^{ME}(\cdot; X, f_R)] = \tfrac{1}{2}\big(P_{X_2}((c, d))\mathbb{1}_{(a,b)}(X_1) - P_{X_1}((a, b))\mathbb{1}_{(c,d)}(X_2)\big)$$

and hence, recalling that $(a, b) \subset \mathcal{X}_1$, $(c, d) \subset \mathcal{X}_2$ and $(a, b) \times (c, d) \subset \mathcal{X}^C$, we have

$$\|\varphi_i[v^{ME}(\cdot; X, f_R)]\|^2_{L^2(\mathbb{P})} = \tfrac{1}{4}\big(P_{X_2}((c, d))^2 P_{X_1}((a, b)) + P_{X_1}((a, b))^2 P_{X_2}((c, d))\big) > 0.$$

Note that $f_R \in H_X$ satisfies $\|f_R\|_{H_X} = \|f_R\|_{L^2(P_X)} = 0$, and hence $f_R \in [0]$-equivalence class of H_X. Since the marginal Shapley formula for 0 and f_R yields different outputs, the operator $(\bar{\mathcal{E}}^{ME}, H_X)$ is ill-posed.

The above discussion motivates us to focus our investigation on the case $\tilde{P}_X \ll P_X$. In this case, the Radon-Nikodym derivative of $\tilde{P}_X$ with respect to P_X exists and encodes information about feature dependencies. The following lemma, which will be helpful for our analysis, provides a representation of the Radon-Nikodym derivative and the space $L^2(\tilde{P}_X)$.

Lemma 3.14. *Suppose $\tilde{P}_X \ll P_X$. Let $r := \frac{d\tilde{P}_X}{dP_X}$. Then $L^2(\tilde{P}_X)$ can be identified with the weighted L^2-space $L^2_r(P_X)$ where*

(3.13)
$$r = \frac{1}{2^n}\sum_{S\subseteq N} r_S \geq \frac{1}{2^{n-1}}, \quad \text{where} \quad 0 \leq r_S := \frac{dP_{X_S}\otimes P_{X_{-S}}}{dP_X} \in L^1(P_X) \quad \text{with} \quad \|r_S\|_{L^1(P_X)} = 1.$$

We next establish the conditions when the marginal game is continuous, that is, bounded. This will help us to determine when the marginal operator on H_X is bounded.

Lemma 3.15 (game boundedness). *Suppose $\tilde{P}_X \ll P_X$. Let r, r_S be as in Lemma 3.14.*

(i) Suppose that $r = \frac{d\tilde{P}_X}{dP_X} \in L^\infty(P_X)$, which is equivalent to

(BG) $$[P_{X_S}\otimes P_{X_{-S}}](A\times B) \leq M\cdot P_X(A\times B),\ A\in\mathcal{B}(\mathbb{R}^{|S|}),\ B\in\mathcal{B}(\mathbb{R}^{|-S|})$$

for any $S\subseteq N$ and some $M\geq 0$. Then the map $f\in H_X \mapsto v^{ME}(S;X,f)\in L^2(\mathbb{P})$, $S\subseteq N$, is bounded.

(ii) Let $\varnothing \neq S\subset N$. Suppose that either

(UG1)
$$\sup\left\{\frac{[P_{X_S}\otimes P_{X_{-S}}](A\times B)}{P_X(A\times B)}\cdot P_{X_{-S}}(B),\ A\in\mathcal{B}(\mathbb{R}^{|S|}),\ B\in\mathcal{B}(\mathbb{R}^{|-S|}), P_X(A\times B)>0\right\} = \infty.$$

or the non-negative, well-defined Borel function

(UG2) $$\rho(x_S) := \int r_S^{1/2}(x_S,x_{-S})P_{X_{-S}}(dx_{-S})$$

with values in $\mathbb{R}\cup\{\infty\}$ is not P_{X_S}-essentially bounded.
Then the map $f\in H_X\mapsto v^{ME}(S;X,f)\in L^2(\mathbb{P})$ is unbounded.

Proof. By Lemma B.1 the condition (BG) is equivalent to $r = \frac{d\tilde{P}_X}{dP_X}\in L^\infty(P_X)$. Then

$$\begin{aligned}\|v(S;X,f)\|^2_{L^2(\mathbb{P})} &\leq \int |f(x)|^2\,[P_{X_S}\otimes P_{X_{-S}}](dx_S,dx_{-S})\\ &= \int r_S|f(x)|^2P_X(dx) \leq 2^n\|r\|_{L^\infty(P_X)}\int |f(x)|^2\,P_X(dx)\end{aligned}$$

for any $S\subseteq N$, and where r_S is given by (3.13). This proves (i).

Let $\varnothing\neq S\subset N$. First, suppose that the condition (UG1) holds. Suppose that $A\in\mathcal{B}(\mathbb{R}^{|S|})$, $B\in\mathcal{B}(\mathbb{R}^{|-S|})$, and $P_X(A\times B)>0$. Set $f(x) = 1_A(x_S)\cdot 1_B(x_{-S})$. Then

$$\begin{aligned}\frac{\mathbb{E}[v^{ME}(S;X,f)]^2}{\|f\|^2_{L^2}(P_X)} &= \frac{1}{P_X(A\times B)}\int\Big(\int 1_A(x_S)\cdot 1_B(x_{-S})\,P_{X_{-S}}(dx_{-S})\Big)^2P_{X_S}(dx_S)\\ &= \frac{P_{X_S}(A)\cdot(P_{X_{-S}}(B))^2}{P_X(A\times B)} = \frac{[P_{X_S}\otimes P_{X_{-S}}](A\times B)}{P_X(A\times B)}\cdot P_{X_{-S}}(B).\end{aligned}$$

Then (UG1) and the relationship above imply that the map $f\in H_X\mapsto v^{ME}(S;X,f)\in L^2(\mathbb{P})$ is unbounded. This proves the first part of (ii).

To prove the second part, suppose the map $f \in H_X \mapsto v^{ME}(S;X,f) \in L^2(\mathbb{P})$ is bounded. Then there exists $c_* > 0$ such that for any $f \in H_X$ we have

$$\int \Big(\int f(x_S, x_{-S}) P_{X_{-S}}(dx_{-S}) \Big)^2 P_{X_S}(dx_S) \le c_* \int f^2(x) P_X(dx).$$

Let $A \in \mathcal{B}(\mathbb{R}^{|S|})$. Then, by above and the definition of $r_S \ge 0$, we obtain

$$\int_A \rho^2(x_S) P_{X_S}(dx_S) = \int_A \Big(\int r_S^{1/2}(x_S, x_{-S}) P_{X_{-S}}(dx_{-S}) \Big)^2 P_{X_S}(dx_S) \le c_* \int 1_A(x_S) P_{X_S}(dx_S).$$

Since A was arbitrary, $0 \le \rho^2 \le c_*$ P_{X_S}-almost surely. This proves the second part of (ii). ■

Theorem 3.16 (game value boundedness). *Let h, X be as in Definition 3.1. Suppose $\tilde{P}_X \ll P_X$, and let r, r_S be as in Lemma 3.14.*

(i) Suppose (BG) holds. Then $H_X = L^2(P_X)$ and for $f \in L^2(P_X)$

$$\begin{aligned} &\|\bar{\mathcal{E}}_i^{ME}[f;h,X]\|_{L^2(\mathbb{P})} \\ &\le \Big(1 + 2 \cdot \max_{S \subseteq N} \|r_S - 1\|_{L^\infty(P_X)}\Big) \Big(\sum_{S \subseteq N \setminus \{i\}} |w(S,n)| \Big) \|f\|_{L^2(P_X)} \end{aligned} \tag{3.14}$$

Consequently, $(\bar{\mathcal{E}}^{ME}, H_X)$ is bounded.

(ii) Suppose there exist two distinct indices $i, j \in N$ such that

$$\text{(UO)} \quad \sup \left\{ \frac{[P_{X_i} \otimes P_{X_j}](A \times B)}{P_{(X_i,X_j)}(A \times B)} \cdot P_{X_j}(B), \;\; A, B \in \mathcal{B}(\mathbb{R}), P_{(X_i,X_j)}(A \times B) > 0 \right\} = \infty.$$

Suppose that the weights in (3.9) satisfy the non-negativity condition (NN) and

$$\sum_{S \subseteq N \setminus \{i,j\}} w(S,n) > 0. \tag{3.15}$$

Then $(\bar{\mathcal{E}}_i^{ME}, H_X)$, $(\bar{\mathcal{E}}_j^{ME}, H_X)$, and $(\bar{\mathcal{E}}^{ME}, H_X)$ are unbounded linear operators.

Proof. By Lemma B.1 the condition (BG) is equivalent to $\tilde{P}_X \ll P_X$ with $\frac{d\tilde{P}_X}{dP_X} \in L^\infty(P_X)$. Thus by Proposition (3.22) we have $H_X = L^2(P_X)$ and hence for every $f \in L^2(P_X)$ we have

$$\|\bar{\mathcal{E}}_i^{ME}[f;h,X]\|_{L^2(\mathbb{P})} \le \|\bar{\mathcal{E}}_i^{ME}[f;h,X] - \bar{\mathcal{E}}_i^{CE}[f;h,X]\|_{L^2(\mathbb{P})} + \|\bar{\mathcal{E}}_i^{CE}[f;h,X]\|_{L^2(\mathbb{P})} =: I_1 + I_2.$$

Combining the bound for I_1 given by Proposition 3.22 and the bound for I_2 obtained from Theorem 3.2(i) together with the definition 3.1, we obtain 3.14. This proves (i).

Suppose next (UO) holds for some distinct $i, j \in N$. Let $w(S,n)$ be the weights as in (3.1). Define

$$w_{\{i,j\}} := \sum_{S \subset N : i \notin S, j \notin S} w(S,n), \quad w_{\{i\}} := \sum_{S \subset N : i \notin S, j \in S} w(S,n), \quad w_{\{j\}} := \sum_{S \subset N : i \in S, j \notin S} w(S,n). \tag{3.16}$$

Suppose (NN) holds, that is, $w(S,n) \geq 0$ for $S \subset N$. Suppose also (3.15) holds for the indices i, j. Then

$$\underline{w}_{i,j} := \min\{|w_{\{i\}}|, |w_{\{j\}}|, |w_{\{i,j\}}|\} > 0. \tag{3.17}$$

For instance, for the Shapley value, one always has $w_{\{i\}} = w_{\{i,j\}} = \frac{1}{2}$ (which allows one to simplify some of the computations below; cf. Remark 3.18).

First, consider a special case $n = 2$. In that case, we have $i = 1, j = 2$. Let $R = A \times B \subseteq \mathbb{R}^2$ where A, B are Borel sets. Denote $f_R := \mathbb{1}_R(x_1, x_2) = \mathbb{1}_A(x_1)\mathbb{1}_B(x_2)$. Then, by (3.1) for $i = 1$, we obtain

$$\begin{aligned}\bar{\mathcal{E}}_1^{ME}[f_R] = h_1[v^{ME}(\cdot; X, f_R)] &= w(\varnothing)\big[v^{ME}(\{1\}; X, f_R) - v^{ME}(\varnothing; X, f_R)\big] \\ &\quad + w(\{2\})\big[v^{ME}(\{1,2\}; X, f_R) - v^{ME}(\{2\}; X, f_R)\big] \\ &= w(\varnothing)\big(\mathbb{1}_A(X_1)P_{X_2}(B) - P_X(R))\big) \\ &\quad + w(\{2\})\big(\mathbb{1}_R(X_1, X_2) - \mathbb{1}_B(X_2)P_{X_1}(A)\big)\end{aligned}$$

where we suppress the dependence on n in the coefficients $w(S, n)$.

Let us denote $p := P_X(R)$, $\alpha := P_{X_1}(A)$, and $\beta := P_{X_2}(B)$. Then

$$\begin{aligned}(\bar{\mathcal{E}}_1^{ME}[f_R])^2 = w^2(\varnothing)\big(\mathbb{1}_A(X_1)\beta^2 + p^2 - \mathbb{1}_A(X_1)2\beta p\big) \\ + w^2(\{2\})\big(\mathbb{1}_R(X_1, X_2) + \mathbb{1}_B(X_2)\alpha^2 - \mathbb{1}_R(X_1, X_2)2\alpha\big) \\ + 2w(\varnothing)w(\{2\})(\mathbb{1}_R(X_1, X_2)\beta - \mathbb{1}_R(X_1, X_2)\alpha\beta - p\mathbb{1}_R(X_1, X_2) + \mathbb{1}_B(X_2)\alpha p).\end{aligned}$$

Then, taking the expectation we obtain

$$\begin{aligned}&\mathbb{E}[(\bar{\mathcal{E}}_1^{ME}[f_R])^2] \\ &= w^2(\varnothing)\big(\alpha\beta^2 + p(p - 2\alpha\beta)\big) + w^2(\{2\})\big(p(1 - 2\alpha) + \beta\alpha^2\big) + 2w(\varnothing)w(\{2\})p(\beta - p) \\ &\geq w^2(\varnothing)\alpha\beta^2 + w^2(\{2\})\beta\alpha^2 - 2p(w^2(\varnothing) + w^2(\{2\})) - 2p|w(\varnothing)w(\{2\})|.\end{aligned}$$

Note that $\|f_R\|^2_{L^2(P_X)} = P_X(R) = p$ and hence, assuming that $p > 0$, we conclude

$$\frac{1}{\|f_R\|^2_{L^2(P_X)}}\mathbb{E}[(\bar{\mathcal{E}}_1^{ME}[f_R])^2] \geq \underline{w}^2\left(\frac{\alpha\beta^2 + \beta\alpha^2}{p}\right) - 6\overline{w}^2$$

where $\underline{w} := \min_{S \subset N} |w(S)|$ and $\overline{w} := \max_{S \subset N} |w(S)|$. Now if (UO) and (3.15) hold for $i = 1$ and $j = 2$, then (3.17) holds for $i = 1$ and $j = 2$ and hence $\underline{w}^2 > 0$ in the inequality above. Then the right-hand side of the inequality is unbounded and, hence, $(\bar{\mathcal{E}}_1^{ME}, H_X)$ is unbounded.

Performing similar calculations for $\bar{\mathcal{E}}_2^{ME}$, we come to the conclusion that if (UO) and (3.15) hold for $i = 1$ and $j = 2$, then $(\bar{\mathcal{E}}_2^{ME}, H_X)$ is unbounded. This proves part (iii) for $n = 2$.

Next, consider a general case of $n \geq 2$. Suppose (UO) holds with for some distinct $i, j \in \{1, 2, \ldots, n\}$. Let $R = A \times B \subseteq \mathbb{R}^2$, where A, B are Borel sets. Define a function of n variables as follows $f_R(x_1, x_2, \ldots, x_n) := \mathbb{1}_R(x_i, x_j)$. By construction, f_R does not depend explicitly on x_k for each $k \in N \setminus \{i, j\}$, and hence by Theorem 3.8(vi), $T := \{i, j\}$ is a carrier

for $v^{ME}(\cdot; X, f_R)$. Hence, by (3.1), we obtain

$$\begin{aligned}\bar{\mathcal{E}}_i^{ME}[f_R] &= w_{\{i,j\}}\Big(v^{ME}(\{i\}; X, f_R) - v^{ME}(\varnothing; X, f_R)\Big) \\ &\quad + w_{\{i\}}\Big(v^{ME}(\{i,j\}; X, f_R) - v^{ME}(\{j\}; X, f_R)\Big) \\ \bar{\mathcal{E}}_j^{ME}[f_R] &= w_{\{i,j\}}\Big(v^{ME}(\{j\}; X, f_R) - v^{ME}(\varnothing; X, f_R)\Big) \\ &\quad + w_{\{j\}}\Big(v^{ME}(\{i,j\}; X, f_R) - v^{ME}(\{i\}; X, f_R)\Big)\end{aligned}$$

where $w_{\{i\}}, w_{\{j\}}$, and $w_{\{i,j\}}$ are defined in (3.16).

Note that for each $S \subseteq T = \{i, j\}$

$$v^{ME}(S; X, f_R) = v^{ME}(S; (X_i, X_j), \mathbb{1}_R(x_i, x_j)).$$

Then, denoting $\alpha := P_{X_i}(A)$, $\beta := P_{X_j}(B)$, $p := P_{(X_i,X_j)}(R)$ and proceeding as in the case $n = 2$, we obtain

$$\frac{\mathbb{E}[(\bar{\mathcal{E}}_i^{ME}[f_R])^2]}{\|f_R\|^2_{L^2(P_X)}}, \frac{\mathbb{E}[(\bar{\mathcal{E}}_j^{ME}[f_R])^2]}{\|f_R\|^2_{L^2(P_X)}} \geq \underline{w}_{i,j}^2\left(\frac{\alpha\beta^2 + \beta\alpha^2}{p}\right) - 6\overline{w}_{i,j}^2, \tag{3.18}$$

where $\underline{w}_{i,j}$ is defined in (3.17), and $\overline{w}_{i,j} := \max\{|w_{\{i\}}|, |w_{\{j\}}|, |w_{\{i,j\}}|\}$, where we have assumed that $\|f_R\|^2_{L^2(P_X)} = P_{(X_i,X_j)}(R) = p > 0$.

By (3.17) the coefficient $\underline{w}_{i,j} > 0$, and hence (UO) for the given distinct indices $i, j \in N$ together with (3.18) imply that $\bar{\mathcal{E}}_i^{ME}$, $\bar{\mathcal{E}}_j^{ME}$, and $\bar{\mathcal{E}}^{ME}$ are unbounded on H_X. This proves *(iii)*.■

Remark 3.17. When $r_S = 1$, $\forall S \subseteq N$ (that is the predictors are independent and hence $\tilde{P}_X = P_X$), the bound in (3.14) becomes that of (3.3).

Remark 3.18. In the supplemental material, we shall explain how condition (UO) for the unboundedness of the marginal operator emerges naturally by considering the case of h being the Shapley value φ for which the weights $w(S, n)$ are known (cf. (2.2)).

Remark 3.19. In Theorem 3.16, we showed that if $r = \frac{d\tilde{P}_X}{dP_X}$ exists and belongs to $L^\infty(P_X)$, then $H_X = L^2(P_X)$. It turns out that the converse is true as well. That is, if $\tilde{P}_X \ll P_X$ and $H_X = L^2(P_X)$, then $r \in L^\infty(P_X)$; see Lemma SM4.

Theorem 3.16 suggests that there are two regimes for well-defined marginal explanations. In the first one, the explanations are bounded but the Lipschitz bound increases as the strength of dependencies increases. In the second one, the marginal operator is unbounded. Below are two examples that illustrate both cases.

Example 3.3. Let $f(x) = \frac{1}{\sqrt{\delta}}(x_1 - x_2)$, $\delta > 0$. Let $X = (X_1, X_2)$ with $\mathbb{E}[X_i] = 0$. Let $X_i = Z + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \delta^2)$, $i \in \{1, 2\}$, where $Z \sim \mathcal{N}(0, 1)$, and $\epsilon_1, \epsilon_2, Z$ are independent.

First, note that $f(X) = \frac{1}{\sqrt{\delta}}(\epsilon_1 - \epsilon_2)$ and hence, by independence of ϵ_1 and ϵ_2, we obtain

$$\|f\|^2_{L^2(P_X)} = Var(f(X)) = \delta^{-1} \cdot (Var(\epsilon_1) + Var(\epsilon_2)) = 2\delta.$$

Then, since $\bar{\mathcal{E}}_1^{ME}[f;\varphi,X] = \frac{1}{\sqrt{\delta}}X_1, \bar{\mathcal{E}}_2^{ME}[f;\varphi,X] = -\frac{1}{\sqrt{\delta}}X_2$, we conclude

$$\|\bar{\mathcal{E}}^{ME}[f;\varphi,X]\|^2_{L^2(\mathbb{P})^2} = \frac{2}{\delta} + O(\delta).$$

Thus, as $\delta \to 0^+$, $\|f\|_{L^2(P_X)} \to 0$, but $\|\bar{\mathcal{E}}^{ME}[f;\varphi,X]\|_{L^2(\mathbb{P})^2} \to \infty$.

Example 3.4. Let $Y \sim \exp(1)$ and $Z \sim \mathcal{N}(0,1)$. Let $X = (X_1, X_2)$ be a random vector with values in $\mathbb{R}^2$ such that $P_X = \frac{1}{2}P_{(Y,Y)} + \frac{1}{2}P_Z \otimes P_Z$. By design, $\tilde{P}_X \ll P_X$ and hence the marginal Shapley value is a well-defined operator on $L^2(P_X)$. Take $t \in \mathbb{R}_+$ and define a square $R^t := [t-1,t] \times [t,t+1] =: I_1^t \times I_2^t$. Then, since $\lim_{t\to+\infty} \frac{(P_Y(I_j^t))^2}{P_Z(I_j^t)} = \infty$, $j \in \{1,2\}$, we have

$$\lim_{t\to+\infty} \frac{[P_{X_1} \otimes P_{X_2}](R^t)}{P_X(R^t)} \cdot P_{X_j}(I_j^t) = \infty, \quad j \in \{1,2\}$$

which by Theorem 3.16 implies that the marginal Shapley value on $L^2(P_X)$ is unbounded.

The absolute continuity condition also allows to express the Wasserstein distance of the two probability measures using the Radon-Nikodym derivative, explaining how the latter controls the strength of dependencies among the predictors.

Lemma 3.20. *Let $X = (X_1, \dots, X_n) \in L^1(\mathbb{P})^n$. Let r, r_S be as in Lemma 3.14.*

$$W_1(\tilde{P}_X, P_X) \le \int |x| \cdot |r(x) - 1| \, P_X(dx) \le \frac{1}{2^n} \sum_{S \subseteq N} \int |x| \cdot |r_S(x) - 1| \, P_X(dx) < \infty \tag{3.19}$$

Proof. Follows from Lemma 3.14, Lemma B.4, and the triangle inequality. ∎

The above lemma illustrates that dependencies are controlled by the Radon-Nikodym derivative. When $r = 1$, then $r_S = 1$, $S \subseteq N$ and the two measures coincide. When r deviates from 1, the dependencies start to impact the distance. As a consequence, the marginal and conditional explanations start to differ from one another. The estimate on this difference is discussed below in the special case when the Radon-Nikodym derivative is bounded.

Lemma 3.21. *Suppose $\tilde{P}_X \ll P_X$. Let r, r_S be as in Lemma 3.14. Suppose $r \in L^2(P_X)$.*
(i) Let $f \in L^2_{r^2}(P_X)$. Then, for every $S \subseteq N$

$$\mathbb{E}\big[(v^{CE}(S;X,f) - v^{ME}(S;X,f))^2\big] \le \|(r_S - 1) \cdot f\|^2_{L^2(P_X)} < \infty. \tag{3.20}$$

(ii) Let $f \in L^2_{r^2}(P_X)$. Let $h[N,v]$ be a game value in the form (3.1). Then

$$\begin{aligned} &\Big(\mathbb{E}\big[\big(h[N,v^{CE}(\cdot;X,f)] - h[N,v^{ME}(\cdot;X,f)]\big)^2\big]\Big)^{1/2} \\ &\le \sum_{S \subseteq N\setminus\{i\}} |w(S,n)| \Big(\|(r_S - 1) \cdot f\|_{L^2(P_X)} + \|(r_{S\cup\{i\}} - 1) \cdot f\|_{L^2(P_X)}\Big). \end{aligned} \tag{3.21}$$

Proof. Take $f \in L^2_{r^2}(P_X)$. Take $S \subset N$. Then, by Lemma 3.14 and Lemma B.2, we have $f \in L^2_{1+r_S^2}(P_X)$ and

$$\begin{aligned}
&\mathbb{E}\big[v^{CE}(S;X,f) - v^{ME}(S;X,f)\big]^2 = \mathbb{E}_{x_S \sim P_{X_S}}\big[\big(\mathbb{E}[f(x_S, X_{-S})|X_S = x_S] - \mathbb{E}[f(x_S, X_{-S})]\big)^2\big] \\
&= \int \Big(\int f(x_S, x_{-S}) P_{X_S|X_{-S}=x_{-S}}(dx_{-S}) - \int f(x_S, x_{-S}) P_{X_{-S}}(dx_{-S}) \Big)^2 P_{X_S}(dx_S) \\
&\leq \|(r_S - 1) \cdot f\|^2_{L^2(P_X)}.
\end{aligned}$$

This proves (i). The item (ii) follows directly from (i) and the representation (3.1) of h. ∎

Proposition 3.22 (approximation). *Let h, X be as in Definition 3.1 and r, r_S as in Lemma 3.14. Suppose $\tilde{P}_X \ll P_X$ with $r = \frac{d\tilde{P}_X}{dP_X} \in L^\infty(P_X)$. Then $H_X = L^2(P_X)$, $r_S \in L^\infty(P_X)$, $S \subseteq N$, and for $f \in L^2(P_X)$*

$$\begin{aligned}
&\|\bar{\mathcal{E}}_i^{CE}[f;h,X] - \bar{\mathcal{E}}_i^{ME}[f;h,X]\|_{L^2(\mathbb{P})} \\
&\leq 2 \cdot \Big(\max_{S \subseteq N} \|r_S - 1\|_{L^\infty(P_X)} \Big) \Big(\sum_{S \subseteq N \setminus \{i\}} |w(S,n)| \Big) \|f\|_{L^2(P_X)}, \ i \in N.
\end{aligned} \tag{3.22}$$

Proof. Let $r \in L^\infty(P_X)$. Then by Proposition 3.13 we have $H_X \cong (L^2(\tilde{P}_X), \|\cdot\|_{L^2(P_X)}) \subseteq L^2(P_X)$. Take $f \in L^2(P_X)$. Then, by the definition of Radon-Nikodym derivative, we obtain

$$\int |f(x)|^2 \tilde{P}_X(dx) = \int r(x)|f(x)|^2 \, P_X(dx) \leq \|r\|_{L^\infty(P_X)} \int |f(x)|^2 \, P_X(dx) < \infty,$$

and hence $f \in H_X$. This proves that $H_X = L^2(P_X)$.

The remaining part of the statement follows from Lemma 3.14 and Lemma 3.21(ii). ∎

Remark 3.23. It is crucial to point out that μ-consistency of explanations is merely a stability (continuity) requirement with the Lipschitz bound determining the relative scale between explanation differences and the differences of associated models. Thus, the three criteria that are useful for the design of explanations are: a) μ-consistency which determines the type of similarity of explanations, b) the Lipschitz bound which determines relative scaling of explanations and models, and c) the game which determines the "shape" of explanations.

Global feature importance. The above analysis extends to global feature attributions inherited from game values as follows. Given a game value h and predictors $X = (X_1, \ldots, X_n)$ define the global conditional and marginal attributions by

$$\beta(v, X, f) := \{\beta_i(v, X, f)\}_{i \in N}, \quad \beta_i := \|h_i[N, v(\cdot; X, f)]\|_{L^2(\mathbb{P})}, \ v \in \{v^{CE}, v^{ME}\}.$$

Then, according to Corollary 3.5 and Theorem 3.8, the global explanations satisfy the continuity condition $|\beta(v, X, f_1 - f_2)| \leq C\|f_1 - f_2\|_{L^2(\mu)}$, $f_1, f_2 \in L^2(\mu)$, with $\mu = P_X$ when $v = v^{CE}$ and $\mu = \tilde{P}_X$ when $v = v^{ME}$ for some $C = C(h)$. Furthermore, if h satisfies conditions of Corollary 3.5(ii) and $v = v^{CE}$, then $C = 1$.

3.3. Splitting of explanation energy on dependencies. We next provide an example that showcases that model's *energy* (in the sense of its squared norm) is split on conditional explanations and some of it is dissipated. To see this, recall that the efficiency property puts a constraint on the vector $\bar{\mathcal{E}}^{CE}[f;h,X]$; its components should add up to $f(X)-\mathbb{E}[f(X)]$. In that case, in light of Corollary 3.5(*ii*), the energy of the conditional explanation vector is bounded by that of the (centered) model:

$$\|\bar{\mathcal{E}}^{CE}[f;h,X]\|^2_{L^2(\mathbb{P})^n}=\sum_{i=1}^n\|\bar{\mathcal{E}}^{CE}_i[f;h,X]\|^2_{L^2(\mathbb{P})}\le\|f-f_0\|^2_{L^2(P_X)},\quad f_0:=\mathbb{E}[f(X)]. \tag{3.23}$$

By contrast, in view of the Rashomon effect [5] and Theorem 3.13, the energy of the model can be significantly lower than that of the marginal explanations.

It is worth mentioning that, when the game value h is efficient, then the independence of explanations (both marginal or conditional) leads to energy conservation. In general, for conditional explanations, we have the following result on the energy conservation.

Lemma 3.24. *Let h be an efficient game value in the form (3.1) with $w(S,n)\ge 0$. The equality in (3.23) is achieved if and only if*

$$\langle f(X)-\bar{\mathcal{E}}^{CE}_i[f;h,X],\bar{\mathcal{E}}^{CE}_i[f;h,X]\rangle_{L^2(\mathbb{P})}=0,\quad\forall i\in N. \tag{3.24}$$

Proof. Following the proof of Corollary 3.5(*ii*) the equality in

$$\|\bar{\mathcal{E}}^{CE}[f]\|^2_{L^2(\mathbb{P})^n}=\sum_{i=1}^n\|\bar{\mathcal{E}}^{CE}_i[f]\|^2_{L^2(\mathbb{P})}\le\sum_{i=1}^n\langle f(X),\bar{\mathcal{E}}^{CE}_i[f]\rangle_{L^2(\mathbb{P})}=\|f\|^2_{L^2(P_X)}$$

is achieved if and only if $\|\bar{\mathcal{E}}^{CE}_i[f]\|^2_{L^2(\mathbb{P})}=\langle f(X),\bar{\mathcal{E}}^{CE}_i[f]\rangle_{L^2(\mathbb{P})}$ for all $i\in N$. ∎

We next show that, under dependencies, model's energy can be dissipated on explanations.

Example 3.5. Let $X=(X_1,\dots,X_n)$ and $f(x)$ be globally Lipschitz. Suppose that $X_i=Z+\epsilon_i$, where $Z\in L^2(\mathbb{P})$ is a latent variable, and $\epsilon_i\sim\mathcal{N}(0,\delta)$, for each $i\in N$.

Define $\bar{f}(x):=\frac{1}{n}\sum_{i=1}^n f(x_i,x_i,\dots,x_i)$. Since $\bar{f}$ is a symmetric function (that is, the order of the input components does not matter) and the Shapley value φ is a symmetric game value, we must have

$$\mathcal{E}^{CE}_i[\bar{f}(X);\varphi,X]=\mathcal{E}^{CE}_j[\bar{f}(X);\varphi,X],\quad i,j\le n.$$

Since f is globally Lipschitz, by the linearity of $\mathcal{E}^{CE}$, we obtain for each $i\in N$

$$\mathcal{E}^{CE}_i[f(X);\varphi,X]=\mathcal{E}^{CE}_i[\bar{f}(X);\varphi,X]+O(\delta)\quad\text{in } L^2(\mathbb{P}).$$

Then, the last two equality imply that for $i,j\le n$

$$\bar{\mathcal{E}}^{CE}_i[f;\varphi,X]=\bar{\mathcal{E}}^{CE}_j[f;\varphi,X]+O(\delta),\quad\text{in } L^2(\mathbb{P}).$$

Hence, by efficiency of φ, for every $j\le n$ we obtain

$$f(X)-f_0=\sum_{i=1}^n\bar{\mathcal{E}}^{CE}_i[f;\varphi,X]=n\cdot\mathcal{E}^{CE}_j[f(X);\varphi,X]+O(\delta)\quad\text{in } L^2(\mathbb{P}),$$

where $f_0:=\mathbb{E}[f(X)]$. This implies that $\|\bar{\mathcal{E}}^{CE}[f;\varphi,X]\|_{L^2(\mathbb{P})^n}=\frac{1}{\sqrt{n}}\|f-f_0\|_{L^2(P_X)}+O(\delta)$.

3.4. Game value extensions to non-cooperative games. In this subsection, we discuss possible extensions of generic linear game values, which are not necessarily in the form (3.1), to non-cooperative games such as marginal and conditional games associated with an ML model. Recall that a cooperative game with n players is a set function v that acts on a finite set of players $N \subset \mathbb{N}$ and satisfies $v(\varnothing) = 0$. Typically, $N = \{1, 2, \dots, n\}$; see Appendix A.

Let V_0 be the set of all cooperative games with finitely many players. Let us next consider set functions that violate the condition $v(\varnothing) = 0$. To this end, let us denote the collection of such games by

$$V = \{(N, v) : \ v(\varnothing) \in \mathbb{R}, \quad v(S) = \tilde{v}(S), \ S \subseteq N, \quad |S| \geq 1, \ \text{for some } (N, \tilde{v}) \in V_0\}. \tag{3.25}$$

One way to construct an extension of a linear game value to V is to incorporate the value $v(\varnothing)$ into the extension itself. In what follows, for each $v \in V$, the cooperative game $\tilde{v}$ denotes its projection onto V_0 as in (3.25) (it agrees with v on non-empty sets). Given a linear game value h, we seek an extension $\bar{h}$ to V that satisfies:

(E1) $\bar{h}[N, \tilde{v}] = h[N, \tilde{v}]$ for $(N, \tilde{v}) \in V_0$,
(E2) $\bar{h}$ is linear on V.

Lemma 3.25. *Let h be a linear game value. An extension $\bar{h}$ satisfying (E1)-(E2) has the representation:*

$$\bar{h}_i[N, v] = h_i[N, \tilde{v}] + \gamma_i v(\varnothing), \quad i \in N = \{1, 2, \dots, n\}, \tag{3.26}$$

where $\{\gamma_i\}_{i=1}^n$ are constants that depend on N. Furthermore, any game in the form (3.26) satisfies properties (E1)-(E2). In addition, if h is symmetric, then $\bar{h}$ is symmetric if and only if $\gamma_i = \gamma_j$, for each $i, j \in N$.

For example, consider the Shapley value φ defined in (2.2). The same formula can be applied to non-cooperative games to construct an extension. In that case, one has $\gamma_i(\varphi, n) = -\frac{1}{n}$ and the extension satisfies $\bar{\varphi}_i[N, v] = \varphi_i[N, \tilde{v}] - \frac{1}{n} v(\varnothing)$. The efficiency property for the extension then reads as $\sum_{i=1}^n \bar{\varphi}_i[N, v] = v(N) - v(\varnothing)$.

Definition 3.26. *Let h be a linear game value and $\bar{h}$ its extension. We say that $\bar{h}$ is centered if $\bar{h}[N, c] = 0$ for any constant non-cooperative game $(N, c) \in V$.*

Notice that the extension of Shapley value we introduced above are centered.

Lemma 3.27. *Let h be a linear game value and $\bar{h}$ its extension with $\gamma = \{\gamma_i\}_{i=1}^n$ as in (3.26). Let u denote a unit, non-cooperative game on N, that is, $u(S) = 1$ for $\forall S \subseteq N$. Then*

(i) $\bar{h}$ is centered if and only if $\gamma = -h[N, \tilde{u}]$.
(ii) $\bar{h}$ is centered if and only if $\bar{h}[N, v] = h[N, (v - v(\varnothing)u)]$.
(iii) If h has the form

$$h_i[N, \tilde{v}] = \sum_{S \subseteq N \setminus \{i\}} w(i, N, S) \big[\tilde{v}(S \cup \{i\}) - \tilde{v}(S)\big], \quad i \in N,$$

where $w(i, N, S)$ $(i \in N, \ S \subseteq N)$ are constants, then it extends by the same formula to a centered game value for non-cooperative games:

$$\bar{h}_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(i, N, S) \big[v(S \cup \{i\}) - v(S)\big], \quad i \in N. \tag{3.27}$$

(*iv*) *Let f be a model and $f_0 = \mathbb{E}[f(X)]$. Then*

$$\bar{h}[N, v(\cdot\,; X, f)] = \bar{h}[N, v(\cdot\,; X, f - f_0)] + f_0 \bar{h}[N, u] \quad \textit{for} \quad v \in \{v^{CE}, v^{ME}\}. \tag{3.28}$$

Hence, if $\bar{h}$ is centered, then $\bar{h}[N, v(\cdot\,; X, f)] = h[N, v(\cdot\,; X, f - f_0)]$, $v \in \{v^{CE}, v^{ME}\}$.

Proof. The proof follows from the linearity of h. ∎

See A in the appendix for more on game values.

Appendix A. Game value axioms.

A cooperative game is a pair (N, v) defined by the finite set of players $N \subset \mathbb{N}$ (typically, $N = \{1, 2, \ldots, n\}$) and a set function v defined on the collection of all subsets $S \subseteq N$, which satisfies $v(\varnothing) = 0$. A set $T \subseteq N$ is called a carrier of v if $v(S) = v(S \cap T)$ for all $S \subseteq N$. A game value is a map $(N, v) \mapsto h[N, v] = (h_i[N, v])_{i \in N}$.

We now list some of useful game value properties:

(LP) (linearity) For two cooperative games (N, v) and (N, w) we have

$$h[N, av + w] = ah[N, v] + h[N, w], \ a \in \mathbb{R}. \tag{A.1}$$

(EP) (efficiency) The sum of the values is equal to the value of the game

$$\sum_{i \in N} h_i[N, v] = v(N). \tag{A.2}$$

(SP) (symmetry) For any permutation π on N and game (N, v)

$$h_{\pi(i)}[N, \pi v] = h_i[N, v], \quad \pi v(\cdot) = v(\pi^{-1}\cdot). \tag{A.3}$$

(NPP) (null player) A null player $i \in N$ is a player that adds no worth to the game v, that is

$$v(S \cup \{i\}) = v(S), \quad S \subseteq N \setminus \{i\}. \tag{A.4}$$

$h[N, v]$ satisfies the null-player property if $h_i[N, v] = 0$ whenever $i \in N$ is a null player.

(TPG) (total payoff growth) There exists strictly increasing $g : \mathbb{R}_+ \times \mathbb{N} \to \mathbb{R}_+$ satisfying $g(0, n) = 0$ and $g(a, n) > 0$ for $a > 0$ such that for all cooperative games (N, v)

$$\textstyle\sum_{i=1}^{n} |h_i[N, v]| \geq g(|v(N)|, |N|) \geq 0. \tag{A.5}$$

(NN) $h[N, v]$ is a game value in the form (3.1) with weights satisfying $w(S, n) \geq 0, S \subset N$.

Appendix B. On probability measures.

Let $\mathcal{B}(\mathbb{R}^k)$ denote the σ-algebra of Borel sets. The space of all Borel probability measures on $\mathbb{R}^k$ is denoted by $\mathscr{P}(\mathbb{R}^k)$. The space of probability measure with finite q-th moment is denoted by

$$\mathscr{P}_q(\mathbb{R}^k) = \Big\{\mu \in \mathscr{P}(\mathbb{R}^k) : \int_{\mathbb{R}^k} |x|^q d\mu(x) < \infty\Big\}.$$

Lemma B.1. *Let μ, ν be probability measures on a measurable space $(\Omega, \mathcal{F})$. Suppose that $\mu \ll \nu$. Then the following statements are equivalent:*

(*i*) $\frac{d\mu}{d\nu} \in L^\infty(\Omega, \mathcal{F}, \nu)$ *in which case* $\mu(A) \leq \left\|\frac{d\mu}{d\nu}\right\|_{L^\infty(\Omega,\mathcal{F},\nu)} \cdot \nu(A)$, $A \in \mathcal{F}$.
(*ii*) *There exists* $M > 0$ *such that* $\mu(A) \leq M \cdot \nu(A)$, *all* $A \in \mathcal{F}$.

Proof. Follows from the definition of the Radon-Nikodym derivative. ∎

Lemma B.2. *Let* $X = (X_1, \ldots, X_n)$, $Z = (Z_1, \ldots, Z_m)$ *be random vectors on a measurable space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and* $P_X \otimes P_Z \ll P_{(X,Z)}$. *Suppose that* $r := \frac{dP_X \otimes P_Z}{dP_{(X,Z)}} \in L^2(P_{(X,Z)})$. *Then for any* $f \in L^2_{1+r^2}(P_{(X,Z)})$, *we have*

$$\int \Big(\int f(x,z) P_X(dx) - \int f(x,z) P_{X|Z=z}(dx) \Big)^2 P_Z(dz) \leq \|(r-1) \cdot f\|^2_{L^2(P_{(X,Z)})}. \tag{B.1}$$

Proof. Take $B \in \mathcal{B}(\mathbb{R}^m)$. Then, by definition of Radon-Nikodym derivative, we have

$$\begin{aligned} &\int_B \Big(\int f(x,z)\, P_X(dx) - \int f(x,z) P_{X|Z=z}(dx) \Big) P_Z(dz) \\ &= \int_B \Big(\int f(x,z)(r(x,z)-1) P_{X|Z=z}(dx) \Big) P_Z(dz). \end{aligned}$$

Since $B \in \mathcal{B}(\mathbb{R}^m)$ is arbitrary, we conclude that for P_Z-almost all z

$$\int f(x,z)\, P_X(dx) - \int f(x,z) P_{X|Z=z}(dx) = \int f(x,z)(r(x,z)-1) P_{X|Z=z}(dx).$$

This implies (B.1). ∎

Definition B.3 **(Wasserstein).** *The Wasserstein distance* W_1 *on* $\mathscr{P}_1(\mathbb{R}^k)$ *is given by [25]*

$$W_1(\mu,\nu) = \sup \Big\{ \int \psi(x)[\mu - \nu](dx), \quad \psi \in Lip_1(\mathbb{R}^k) = \big\{ u : |u(x) - u(x')| \leq |x - x'| \big\} \Big\}.$$

Lemma B.4 **(Wasserstein bound).** *Let* $\mu, \nu \in \mathscr{P}_1(\mathbb{R}^k)$. *Suppose* $\mu \ll \nu$. *Then*

$$W_1(\mu,\nu) \leq \int |x| \cdot |r(x) - 1|\, \nu(dx) < \infty, \quad r := \frac{d\mu}{d\nu}. \tag{B.2}$$

Proof. Take $\psi \in Lip_1(\mathbb{R}^k)$. Then, by Definition B.3 we have

$$W_1(\mu,\nu) = \sup \Big\{ \int (\psi(x) - \psi(0))(r-1)\, \nu(dx),\ \psi \in Lip_1(\mathbb{R}^k) \Big\}.$$

Since $\psi \in Lip_1(\mathbb{R}^k)$, $|\psi(x) - \psi(0)| \leq |x|$, which implies (B.2). ∎

References.

[1] K. Aas, M. Jullum, and A. Løland, Explaining individual predictions when features are dependent more accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021.
[2] K. Aas, T. Nagler, M. Jullum, A. Løland, Explaining predictive models using Shapley values and non-parametric vine copulas, *Dependence modeling 9*, (2021), 62-81.
[3] D. Alvarez-Melis and T. S. Jaakkola, Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS18, pp. 77867795, Red Hook, NY, USA, 2018. Curran Associates Inc.

[4] J. F. Banzhaf, Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review* 19, 317–343, (1965).

[5] L. Breiman, Statistical Modeling: The two cultures. *Stat. Science, 16-3, 199-231*, (2001).

[6] I. Covert, S. Lundberg, S.-I. Lee, Explaining by Removing: A Unified Framework for Model Explanation. *arXiv preprint arXiv:2011.14878v2*, (2022).

[7] H. Chen, S. Lundberg, and S.-I. Lee. Explaining models by propagating Shapley values of local components. *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, pages 261–270, 2021.

[8] H. Chen, J. Danizek, S. Lundberg, S.-I. Lee, True to the Model or True to the Data. *arXiv preprint arXiv:2006.1623v1*, (2020).

[9] J. Chen, L. Song, M. J. Wainwright, Mi. I. Jordan, L-Shapley and C-Shapley: an efficient model interpretation for structured data. In *7th international conference on Learning representation, New Orleans, USA (2019b)*.

[10] Shapley-based Explainable AI for Clustering Applications in Fault Diagnosis and Prognosis, *arXiv preprint arXiv:2303.14581*, (2023).

[11] T. W. Campbell, H. Roder, R. W. Georgantas III, and J. Roder. Exact Shapley values for local and model-true explanations of decision tree ensembles. *Machine Learning with Applications*, page 100345, 2022.

[12] Equal Credit Opportunity Act (ECOA), https://www.fdic.gov/regulations/laws/rules/6000-1200.html.

[13] R. Elshawi, M. H. Al-Mallah and S. Sakr, On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making 19*, No. 146 (2019).

[14] D. C. Elton, Self-explaining AI as an alternative to interpretable AI, *arXiv preprint arXiv:2002.05149v6*, (2020).

[15] K. Filom, A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, On marginal feature attributions of tree-based models, *Foundations of Data Science, AIMS*, DOI: 10.3934/fods.2024021, (early access, 2024).

[16] A. Fisher, C. Rudin, F. Dominici All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research 20 (2019)*, (2019).

[17] J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, Vol. 29, No. 5, 1189-1232, (2001).

[18] P. Hall, N. Gill, *An Introduction to Machine Learning Interpretability*, O'Reilly. (2018).

[19] T. Hastie, R. Tibshirani and J. Friedman *The Elements of Statistical Learning*, 2-nd ed., Springer series in Statistics (2016).

[20] L. Hu, J. Chen, V. N. Nair and A. Sudjianto, Locally interpretable models and effects based on supervised partitioning (LIME-SUP), *Corporate Model Risk*, Wells Fargo, USA (2018).

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

[22] H. Ji, K. Lafata, Y. Mowery, D. Brizel, A. L. Bertozzi, F.-F. Yin, C. Wang, Post-Radiotherapy PET Image Outcome Prediction by Deep Learning Under Biological Model Guidance: A Feasibility Study of Oropharyngeal Cancer Application *arXiv preprint*, (2021).

[23] M. Jullum, A. Redelmeier, K. Aas, Efficient and simple prediction explanations with groupShapley: a practical perspective, *XAI.it 2021-Italian Workshop on explainable artificial intelligence*.

[24] Y. Kamijo, A two-step Shapley value in a cooperative game with a coalition structure. *International Game Theory Review*, 11 (2), 207–214.

[25] L.V. Kantorovich, G. Rubinstein On a space of completely additive functions, *Vestnik Leningradskogo Universiteta*, 13 (7), 52–59, (1958).

[26] K. Kotsiopoulos, A. Miroshnikov, K. Filom, A. Ravi Kannan Approximation of group explainers with coalition structure using Monte Carlo sampling on the product space of coalitions and features *arXiv preprint arXiv:2303.10216v1*, (2023).

[27] E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. A. Friedler, Problems with Shapley-value-based explanations as feature importance measures. *arXiv preprint arXiv:2002.11097v2*, (2020).

[28] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Interpretable & Explorable Approximations of Black Box Models. *arXiv preprint arXiv:1707.01154*, (2017).

[29] S. M. Lundberg, G. G. Erion and S.-I. Lee, Consistent individualized feature attribution for tree ensembles, *arXiv preprint arXiv:1802.03888*, (2019).
[30] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, *31st Conference on Neural Information Processing Systems*, (2017).
[31] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1):56–67, 2020.
[32] A. Miroshnikov, K. Kotsiopoulos, K. Filom and A. Ravi Kannan, Stability theory of game-theoretic group feature explanations for machine learning models. *arXiv preprint arXiv:2102.10878v5*, (2024).
[33] A. Miroshnikov, K. Kotsiopoulos, R. Franks and A. Ravi Kannan, Wasserstein-based fairness interpretability framework for machine learning models, *Machine Learning*, 1–51, Springer, (2022).
[34] L. H. B. Olsen, I. K. Glad, M. Jullum, K. Aas, Using Shapley Values and Variational Autoencoders to Explain Predictive Models with Dependent Mixed Features, *Journal of Machine Learning Research*, 23(213):1-51, (2022)
[35] G. Owen, Values of games with a priori unions. *In: Essays in Mathematical Economics and Game Theory (R. Henn and O. Moeschlin, eds.)*, Springer, 76–88 (1977).
[36] J. Pearl, Causality. *Cambridge University Press*, (2000).
[37] Y. A. Reshef, D.N. Reshef, H. K. Finucane, P. C. Sabeti, M. Mitzenmacher, Measuring dependence powerfully and equitably. *Journal of Machine Learning Research*, 17, 1-63 (2016).
[38] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, *22nd Conference on Knowledge Discovery and Data Mining*, (2016).
[39] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
[40] H. L. Royden, P. M. Fitzpatrick, *Real analysis*. Boston: Prentice Hall, 4th ed. (2010).
[41] A. Saabas. treeinterpreter python package https://github.com/andosa/treeinterpreter, 2019.
[42] L. S. Shapley, A value for n-person games, *Annals of Mathematics Studies*, No. 28, 307-317 (1953).
[43] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41, 3, 647-665, (2014).
[44] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, (2017).
[45] M. Sundararajan, A. Najmi, The Many Shapley Values for Model Explanation, *International conference on machine learning*, pages 9269–9278, PMLR, (2020).
[46] J. Teneggi, A. Luster, and J. Sulam, Fast Hierarchical Games for Image Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 45, Issue: 4, 01 April 2023.
[47] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen and V. N. Nair, Explainable Neural Networks based on additive index models *Corporate Model Risk, Wells Fargo, USA, arXiv:1806.01933v1*, (2018).
[48] J. Wang, J. Wiens, S. Lundberg Shapley Flow: A Graph-based Approach to Interpreting Model Predictions *arXiv preprint arXiv:2010.14592*, (2020).
[49] Q. Zhao, T. Hastie, Causal Interpretations of Black-Box Models, *J.Bus. Econ. Stat.*, DOI:10.1080/07350015.2019.1624293, (2019).