Feature grouping as a stabilization mechanism for game-theoretic explanations of machine learning models

Alexey Miroshnikov [*,†] Konstandinos Kotsiopoulos [*,‡] Khashayar Filom[*,§]
Arjun Ravi Kannan[*,¶]

May 17, 2025

Abstract

A well-established and effective approach for explaining machine learning models locally is to utilize game-theoretic feature attributions. Such attributions often stem from a linear game value applied to the conditional and marginal games associated with the predictors and the model at hand. Conditional explanations, which are often infeasible to compute, are known to be consistent with model predictions, while marginal ones explain the model itself. These two types of attributions coincide only when features are independent. Furthermore, when feature dependencies are strong, they may differ drastically: conditional attributions remain stable in $L^2(P_X)$, while marginal ones do not, as shown by the stability theory established in our companion paper [45].

In this paper, we provide a framework for unifying conditional and marginal attributions through feature grouping based on the strength of dependencies. More precisely, we show analytically and verify numerically that grouping features in this way has a stabilizing effect on the marginal operator at both group and individual levels, and enables the approximation of conditional explanations by marginal ones.

keywords: ML interpretability, explanation operator, coalitional game value, Radon-Nikodym derivative, mutual information.

AMS subject classification: 91A06, 91A12, 91A80, 46N30, 46N99, 68T01.

1 Introduction

Modern Machine Learning (ML) modeling algorithms have surpassed traditional statistical techniques, primarily because of their enhanced performance capabilities. Specifically, today's ML models have intricate architectures that boost their predictive power and can process a higher number of inputs. This structural complexity, however, presents certain challenges, notably in their interpretability[1], which necessitate careful consideration in order to address potential concerns, such as a model's trustworthiness.

The term "model explainability" refers to the set of methodologies that evaluate the contribution of each input feature (or predictor) to the model output. These are usually categorized as either post-hoc or self-interpretable. Many post-hoc explanations depend solely on the input and output values in order to generate model explanations, disregarding the internal process that led to the model output itself. These specific techniques are called model-agnostic. On the other hand, self-interpretable methods rely on the internal model architecture to produce explanations, making them model-specific techniques. In general,

[*] Emerging Capabilities Research Group, Discover Financial Services Inc., Riverwoods, IL

[†] first & corresponding author, amiroshn@terpmail.umd.edu, 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 terms interpretability (interpretation) and explainability (explanation) interchangeably. However, the methods discussed in this paper primarily deal with post-hoc explanations derived from a model's results; for details on interpretable models versus post-hoc explanations, see [26].

other explanation methods can also rely on the implementation of the model, the algorithmic steps that led to its construction, and so on.

Explaining an ML model output can be crucial in several industries. For example, in the financial services industry, laws and regulations such as the Equal Credit Opportunity Act [18], requires lenders to inform declined applicants of the main factors that contributed to the adverse decision. In the field of medicine, ML models are used to predict the likelihood of a certain disease or a medical condition, or the result of a medical treatment [33, 61]. In both fields, model explainers can be utilized to assess which inputs likely played a significant role in the prediction and/or decision.

There is a wide array of research that outlines approaches for applying post-hoc as well as self-interpretable methods. There are global methods such as [23, 37] which evaluate the overall contribution of features within the given population, as well as local methods such as the rule-based method [52], locally-interpretable methods [51, 31], and methods such as [57, 41, 13] which provide individualized feature attributions (for a single input) based on the game-theoretic work of Shapley [56]. Other model-specific approaches are the works on explainable neural networks [63] and self-explainable models [6, 19].

Many contemporary interpretability techniques leverage cooperative game theory to generate explanations. These methods calculate a game value in a machine learning context, where the model's prediction serves as the "payout" and features function as "players" contributing to this payout [57, 40, 65, 13, 43, 16, 61, 55, 14, 21]. Some notable game values, such as the Shapley and Owen values [56, 46], are especially valuable in machine learning due to their properties of additivity, efficiency and symmetry (see §A).

Formally, a cooperative game is defined by a set function v over a set of players $N = \{1, \dots, n\}$, where $v(S)$ denotes the payoff that a coalition of players $S \subseteq N$ achieves by cooperating. A game value is a mapping $(N, v) \mapsto h[N, v] \in \mathbb{R}^n$ that assigns to each player an individual contribution to the total payoff $v(N)$.

In our setting, the features $X = (X_1, \dots, X_n)$ are treated as n players in a game $v(\cdot, x; X, f)$ defined by an observation $x \sim P_X$, the random vector X, and a predictive model f. Two notable games introduced in the ML literature [57, 41] – and the focus of our work – are the conditional and marginal games, defined as follows:

$$v_*^{\scriptscriptstyle CE}(S; x, X, f) = \mathbb{E}[f(X)|X_S = x_S] \quad \text{and} \quad v_*^{\scriptscriptstyle ME}(S; x, X, f) = E[f(x_S, X_{-S})].$$

Replacing the fixed observation x with the random vector X makes these games random, and they coincide when the features are independent. However, under feature dependencies, the two games can diverge significantly.

In practice, it is often infeasible to compute the conditional game and corresponding game values such as the conditional Shapley value when feature dependencies are present (see [2] for an approximation using non-parametric vine copulas). This makes the marginal game the only viable option between the two in order to compute explanations in a practical setting. Computing marginal game values, such as the Shapley or Owen value, is still computationally challenging, although efficient algorithms exist for specific model classes [20, 42].

In industry settings like credit scoring and healthcare, explanations need to be interpreted in the context of the underlying game. Heuristic approaches for interpreting the meaning of Shapley values have been proposed in [58, 32, 12]. Broadly speaking, the conditional Shapley value explains observations of $f(X)$, that is, predictions of the model f at inputs $x \sim P_X$, while the marginal Shapley value reflects how the function $f(x)$ transforms individual input values. Notably, [12] refers to conditional explanations as consistent with the data (or *true-to-the-data*), and marginal ones as consistent with the model (or *true-to-the-model*).

To formally distinguish the two games while avoiding heuristics, and to gain a deeper insight into the properties of explanations, we introduced a functional-analytic framework in our companion paper [45]. This framework views explanations as linear maps on appropriate spaces, enabling a stability analysis that rigorously defines consistency in the context of stability.

We show that the distance between the conditional Shapley values of two distinct models is bounded by the distance between their predictions, with a Lipschitz constant of one. The same holds when computing the distance between predictions and the response variable and their corresponding conditional explanations [2]. Thus, the model representation does not influence the explanations. Consequently, consistency – i.e.,

[2] It is possible to define a conditional game directly for the response variable rather than for a model

being true-to-the-data – can be interpreted as the stability of additive explanations on the space of models in $L^2(P_X)$, with a Lipschitz constant of one.

However, this result does not apply to the marginal Shapley value. As feature dependencies strengthen, [45] shows that marginal explanations diverge from the conditional ones, a phenomenon attributed to the Rashomon effect [10], in which the number of distinct models representing the same data increases with stronger dependencies. In this case, the Lipschitz constant for the stability bound of the marginal Shapley values grows and may become infinite, rendering the marginal Shapley value an unbounded (i.e., unstable) map. This makes it unreliable for both computing marginal explanations and approximating conditional ones.

Our contributions. Motivated by the stability analysis in our companion paper [45], in this article we investigate the stabilizing effect of grouping features by dependencies on marginal game values - such as the Shapley value – using the functional-analytic framework from [45]. We demonstrate that marginal game explainers for feature groups - based on the quotient game or coalitional game values, such as the Owen value [46] - allow marginal explanations to approximate conditional ones, while simultaneously enhancing their stability. In other words, grouping by dependencies reduces the instability of marginal explanations in $L^2(P_X)$. Our rigorous analysis validates the observations made in [1], namely, that grouping features by dependencies (e.g. correlation) yields explanations consistent with observed data, while also elucidating the precise mechanisms that drive stability and consistency, and how dependencies influence that behavior.

To practically construct feature groups, we propose a variable hierarchical clustering algorithm that enables the formation of nested groups of predictors based on dependencies rather than simple correlations. This flexible approach effectively reduces the number of explainable components, which varies depending on the chosen dependency threshold applied to the resulting clustering tree. For clustering, we employ a regularized version of mutual information introduced in [50]. This method allows for the construction of nested partitions $\mathcal{P}^\alpha = \{S_1^{(\alpha)}, \dots, S_{m_\alpha}^\alpha\}$ of predictor indices $N = \{1, \dots, n\}$, based on the strength of dependencies $\alpha \in [0,1]$ among features. We utilize this approach on numerical examples that illustrate the stabilizing effect of grouping, where we evaluate the marginal Owen values for tree-based ensembles with symmetric trees, as introduced in [20].

Designing explainers based on predictor groups was previously explored in [1], where groups are formed based on linear dependencies (correlations) rather than a more general form of dependence. The authors of [1] observed that grouping by correlations (and then summing marginal Shapley values) reduces inconsistencies between the marginal and conditional approaches. In the work of [34], the authors focus on quotient game explainers, offering a practical perspective on their implementation, where the groups are formed by domain knowledge rather than statistical dependencies and the conditional game is approximated using method outlined in [2]. These works mainly focus on the Shapley value and investigate practical aspects of grouping. Motivated by these articles, our work focuses on the rigorous analysis of group explainers and the impact of dependencies on their properties, which we believe complements these prior efforts.

To the best of our knowledge, no prior work has provided a rigorous treatment of explanation methods within a functional-analytic framework—one that formally distinguishes between the notions of true-to-the-model and true-to-the-data. Our analysis provides a technical language for understanding the mathematical properties of the underlying games. Moreover, the proposed hierarchical clustering algorithm, when used in conjunction with algorithms such as [20], yields a practical and theoretically grounded approach for computing group-based explanations that enhance stability and bridge the gap between marginal and conditional approaches.

Structure of the paper. In §2, we introduce the requisite concepts, such as the conditional and marginal games, quotient games, the notion of a game value, and in particular, the Shapley value. The properties of conditional and marginal game operators based on quotient games are discussed in §3. Section 4 outlines variable hierarchical clustering using the maximal information coefficient and its application to a synthetic dataset with dependencies. Next, in §5, we present examples that illustrate the theoretical aspects outlined in §3 on both synthetic and real-world data. The paper concludes with an appendix containing all technical proofs.

2 Preliminaries

2.1 Notation and hypotheses

Throughout this article, we consider the joint distribution (X, Y), where $X = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ 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 $\mathcal{X} := \text{supp}(P_X)$. Similarly, we define the measure P_{X_i} with support $\mathcal{X}_i := \text{supp}(P_{X_i})$, $i \in \{1, 2, \dots, n\}$.

Let $S \subseteq N = \{1, 2, \dots, n\}$. 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 $P_{X_S} = \prod_{i \in S} \otimes P_{X_i}$. Let $\mathcal{P} = \{S_1, S_2, \dots, S_m\}$ be a partition of N. We say that the group predictors $X_{S_1}, X_{S_2}, \dots, X_{S_m}$ are independent if $P_X = \prod_{j=1}^{m} \otimes P_{X_{S_j}}$. Furthermore, we define the probability measure $\tilde{P}_X$ as the convex combination of all product measures between X_S and X_{-S}, for all possible $S \subseteq N$, where $-S := N \setminus S$. Specifically, we define

$$\tilde{P}_X := \frac{1}{2^n} \sum_{S \subseteq N} P_{X_S} \otimes P_{X_{-S}} \tag{2.1}$$

where we use the convention that $P_{X_\varnothing} \otimes P_{X_N} = P_{X_N} \otimes P_{X_\varnothing} = P_X$.

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 [22] 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 $\mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$.

Let $X = (X_1, \dots, X_n)$ and $Z = (Z_1, \dots, Z_m)$ be random vectors. Let $D(\cdot, \cdot)$ be a metric on the space of Borel probability measures $\mathscr{P}_k(R^{m+n})$ with k-th finite moment, for some $k \geq 0$. We say that X and Z are (D, ϵ)-weakly independent if $D(P_{(X,Z)}, P_X \otimes P_Z) \leq \epsilon$.

2.2 Explainability and game theory

The objective of a local model explainer $E(x; f, X) = (E_1, \dots, E_n)$ is to quantify the contribution of each predictor X_i, $i \in N$, 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 use ideas from cooperative game theory to construct explainers. A cooperative game with n players is a set function v that acts on a set of size n, say $N = \{1, 2, \dots, n\}$, 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. More details on the game values can be found in §A.

In the ML setting, the features $X = (X_1, X_2, \dots, X_n)$ 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

$$v_*^{\scriptscriptstyle CE}(S; x, X, f) = \mathbb{E}[f(X)|X_S = x_S], \quad v_*^{\scriptscriptstyle ME}(S; x, X, f) = \mathbb{E}[f(x_S, X_{-S})], \tag{2.2}$$

where $v_*^{\scriptscriptstyle CE}(\varnothing; x, X, f) = v_*^{\scriptscriptstyle ME}(\varnothing; x, X, f) := \mathbb{E}[f(X)]$. These are introduced in [57, 41] in the context of the Shapley value [56]

$$\varphi_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n - s - 1)!}{n!} [v(S \cup \{i\}) - v(S)], \quad s = |S|, \, n = |N|. \tag{2.3}$$

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 (2.2) are not cooperative, as they do not satisfy $v(\varnothing) = 0$. In this case, the efficiency property takes the form:

$$\sum_{i=1}^{n} \varphi_i[N,v] = v(N) - v(\varnothing) = f(x) - \mathbb{E}[f(X)], \quad v \in \{v_*^{\scriptscriptstyle CE}(\cdot; x, X, f), v_*^{\scriptscriptstyle ME}(\cdot; x, X, f)\}.$$

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

$$v^{\scriptscriptstyle CE}(S; X, f) = \mathbb{E}[f(X)|X_S], \quad v^{\scriptscriptstyle ME}(S; X, f) = \mathbb{E}[f(x_S, X_{-S})]\big|_{x_s = X_S} \tag{2.4}$$

For these games, the corresponding Shapley values $\varphi[N, v^{\scriptscriptstyle CE}]$ and $\varphi[N, v^{\scriptscriptstyle ME}]$, respectively, are random vectors in $\mathbb{R}^n$. These are well defined (as operators) when the model f belongs to the functional spaces introduced in our companion work [45] and also discussed in Section 3 in the context of quotient games.

Finally, we define a conditional game associated with random variables $Z \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ as follows:

$$\nu_*^{\scriptscriptstyle CE}(S; x, X, Z) = \mathbb{E}[Z|X_S = x_S], \quad \nu^{\scriptscriptstyle CE}(S; X, Z) = \mathbb{E}[Z|X_S], \quad S \subseteq N, \tag{2.5}$$

in which case, for any Borel model f, we have $\varphi[N, v^{\scriptscriptstyle CE}(\cdot; X, f)] = \varphi[N, \nu^{\scriptscriptstyle CE}(\cdot; X, f(X))]$.

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

Given a game value $h[N, v]$, setting $E(X; f, X) = h[N, v(\cdot; X, f)], v \in \{v^{\scriptscriptstyle CE}, v^{\scriptscriptstyle ME}\}$ yields conditional and marginal random model explanations, respectively. In the companion work [45], we studied the boundedness of these (single feature) explanations motivated by the notion of "P_X-consistency" (see [45, p.6]). It turns out that in some cases marginal game-value explanations are unbounded (or unstable). In this article, we provide a remedy based on grouping features, which requires a notion of the quotient game.

Definition 2.2. *Given a cooperative game (N, v) with $N = \{1, 2, \dots, n\}$ and a partition $\mathcal{P} = \{S_1, S_2, \dots, S_m\}$ of N, the quotient game $(M, v^{\mathcal{P}})$, where $M := \{1, 2, \dots, m\}$, is defined by*

$$v^{\mathcal{P}}(A) := v\big(\cup_{j \in A} S_j\big), \qquad A \subseteq M. \tag{2.6}$$

For non-cooperative games, we adapt the same definition; note that one always has $v^{\mathcal{P}}(\varnothing) = v(\varnothing)$.

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

2.3 Motivational example

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 is a wide array of models defined on $\mathcal{X} = \widetilde{\mathcal{X}} = \mathbb{R}^3$ that can represent the same data in an $L^2(P_X)$-sense [10]. For instance, consider $f_\alpha(x) = (1 + \alpha)x_1 + (1 - \alpha)x_2 + x_3$, where $\alpha \in [-\delta^{-\gamma}, \delta^{-\gamma}]$ and $\gamma \in (0, 1)$; in this case the response variable can be expressed by

$$Y = f_\alpha(X) - \alpha(\epsilon_1 - \epsilon_2), \quad \text{where} \quad \|\alpha(\epsilon_1 - \epsilon_2)\|^2_{L^2(\mathbb{P})} = 2\alpha^2\delta^2 \le 2\delta^{2(1-\gamma)}.$$

Thus, the distance between predictions of two distinct models satisfies:

$$\|f_\alpha(X) - f_\beta(X)\|^2_{L^2(\mathbb{P})} = 2(\alpha - \beta)^2\delta^2 \le 8\delta^{2(1-\gamma)}. \tag{2.7}$$

For simplicity of notation in this example, we suppress the dependence on X and write

$$\varphi^{CE}(x;Z) := \varphi[N, \nu_*^{CE}(\cdot;x,X,Z)], \quad \bar{\varphi}^{CE}(x;f) := \varphi[N, v_*^{CE}(\cdot;x,X,f)], \quad \bar{\varphi}^{ME}(x;f) := \varphi[N, v_*^{CE}(\cdot;x,X,f)],$$

for a random variable Z and a Borel model f.

Next, we compute the conditional Shapley explanations for the response variable Y, which yields

$$\varphi_1^{CE}(X;Y) = X_1 + \frac{(\epsilon_1 - \epsilon_2)}{1+\delta^2}, \quad \varphi_2^{CE}(X;Y) = X_2 + \frac{(\epsilon_2 - \epsilon_1)}{1+\delta^2}, \quad \varphi_3^{CE}(X;Y) = X_3, \tag{2.8}$$

where we used the Shapley formula (2.3) with $N = \{1,2,3\}$.

Similarly, the conditional Shapley explanations of $f_\alpha(X)$ are given by

$$\begin{aligned}
\varphi_1^{CE}(X;f_\alpha(X)) &= \varphi_1^{CE}(X;Y) + \frac{\alpha}{2}\left(\epsilon_1 - \epsilon_2 + (X_1+X_2)\frac{\delta^2}{1+\delta^2}\right) = X_1 + O(\delta^{1-\gamma}),\\
\varphi_2^{CE}(X;f_\alpha(X)) &= \varphi_2^{CE}(X;Y) + \frac{\alpha}{2}\left(\epsilon_1 - \epsilon_2 - (X_1+X_2)\frac{\delta^2}{1+\delta^2}\right) = X_2 + O(\delta^{1-\gamma}),\\
\varphi_3^{CE}(X;f_\alpha(X)) &= \varphi_3^{CE}(X;Y) = X_3,
\end{aligned} \tag{2.9}$$

where the term $O(\delta)$ is understood in the $L^2(\mathbb{P})$-sense.

In view of (2.9), as $\delta \to 0^+$, the dependency between X_1 and X_2 strengthens, and the predictions of the models f_α, as well as their conditional explanations, become increasingly similar to those of the response Y, respectively. This fact can be precisely expressed by relating the distances between explanations with those between the predictions and the response, yielding the following stability relationship:

$$\|\varphi^{CE}(X;f_\alpha(X)) - \varphi^{CE}(X;Y)\|^2_{L^2(\mathbb{P})} = \frac{1}{2}\left(1 + \frac{\delta^2}{1+\delta^2}\right) \cdot \|f_\alpha(X) - Y\|^2_{L^2(\mathbb{P})}. \tag{2.10}$$

(2.10) implies that the distance between predictions controls the distance between explanations with the Lipschitz constant that is less than or equal to 1, regardless of the strength of the dependencies. Hence, as long as the predictions of those models are close to the observations of the response $Y = f_0(X)$, their conditional explanations will closely approximate the conditional explanations of the response variable independent of the functional form of the models. Since $\bar{\varphi}^{CE}(X;f_\alpha) = \varphi^{CE}(x;f_\alpha(X))$, (2.10) implies the conditional Shapley values are independent of the representative model up to small additive noise and agree with the explanations of the response variable. This supports the heuristic notion [12] that conditional Shapley explanations $\bar{\varphi}^{CE}(X;f_\alpha)$ are consistent with observations – that is, true-to-the-data.

On the other hand, computing the marginal expectations of f_α, we obtain

$$\bar{\varphi}_1^{ME}(X;f_\alpha) = (1+\alpha)X_1, \quad \bar{\varphi}_2^{ME}(X;f_\alpha) = (1-\alpha)X_2, \quad \bar{\varphi}_3^{ME}(X;f_\alpha) = X_3, \tag{2.11}$$

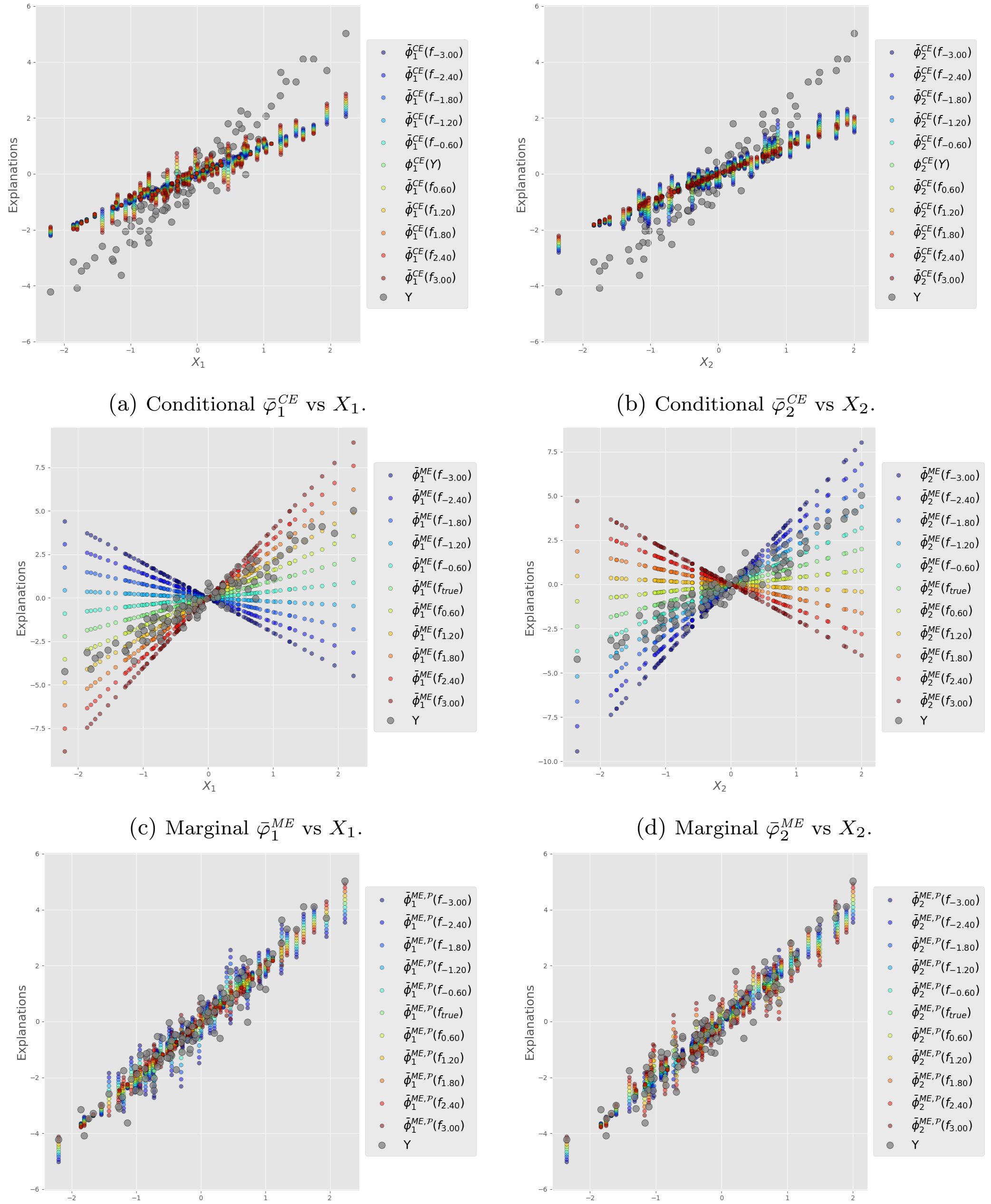
which demonstrates that the marginal explanations depend on the functional form of the model f_α.

Relating the distances between predictions with those of the marginal explanations, we obtain

$$(1+\delta^2)(\alpha-\beta)^2 = \|\bar{\varphi}^{ME}(X;f_\alpha) - \bar{\varphi}^{ME}(X;f_\beta)\|^2_{L^2(\mathbb{P})} = \frac{1}{2}\left(\frac{1+\delta^2}{\delta^2}\right) \cdot \|f_\alpha(X) - f_\beta(X)\|^2_{L^2(\mathbb{P})}, \tag{2.12}$$

where the Lipschitz constant tends to infinity as $\delta \to 0^+$. Thus, as the strength of dependencies increases the two models that have similar predictions are no longer guaranteed to have similar marginal explanations. Moreover, whenever $|\alpha - \beta| \geq \delta^{-\gamma}$ the lower bound in (2.12) satisfies $(1+\delta^2)(\alpha-\beta)^2 \geq \frac{1}{\delta^{2\gamma}} \to \infty$ as $\delta \to 0^+$. The only way to make the marginal explanations of f_α approach to those of f_β is to make sure that $\alpha \to \beta$, in which case, $f_\alpha \to f_\beta$ pointwise. This supports the heuristic notion [12] that marginal Shapley explanations $\bar{\varphi}^{ME}(X;f_\alpha)$ are consistent with the model – that is, true-to-the-model.

Figures 1a-1d show (the single feature) conditional and marginal explanations for models $f_\alpha, \alpha \in [-3,3]$, respectively, compared to the observations of Y.



(a) Conditional $\bar{\varphi}_1^{CE}$ vs X_1.

(b) Conditional $\bar{\varphi}_2^{CE}$ vs X_2.

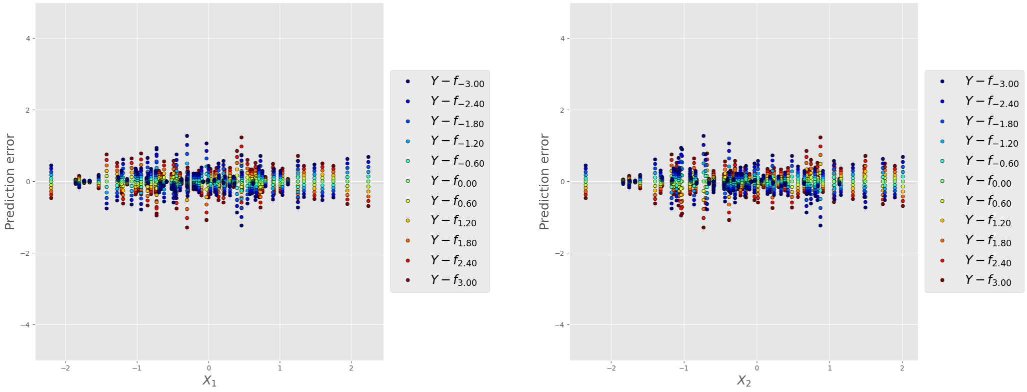
(c) Marginal $\bar{\varphi}_1^{ME}$ vs X_1.

(d) Marginal $\bar{\varphi}_2^{ME}$ vs X_2.

(e) Quotient marginal $\bar{\varphi}_1^{ME,\mathcal{P}}$ vs X_1.

(f) Quotient marginal $\bar{\varphi}_1^{ME,\mathcal{P}}$ vs X_2.

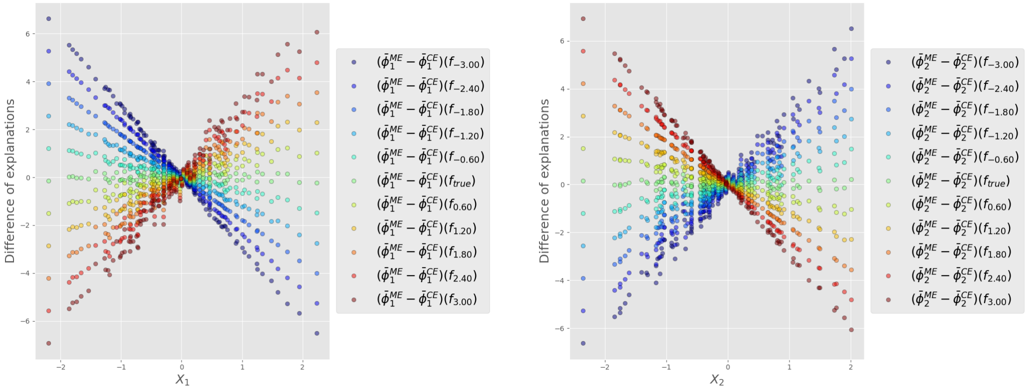
Figure 1: Marginal and conditional explanations.



(a) Prediction error vs X_1.

(b) Prediction error vs X_2.

Figure 2: Difference between the response and model predictions.



(a) $\bar{\phi}_1^{ME} - \bar{\phi}_1^{CE}$ vs X_1.

(b) $\bar{\phi}_2^{ME} - \bar{\phi}_2^{CE}$ vs X_2.

Figure 3: Difference between marginal and conditional explanations.

In this example, when δ is fixed, the Lipschitz bound in (2.12) may be large, but still finite. It is worth noting that there is another regime in which the marginal explanations can be unstable – that is, the Lipschitz bound does not exist (see the companion paper [45] as well as Example 3.1). In that case, it is possible to find models with arbitrarily close predictions, but arbitrarily different marginal explanations.

Finally, computing the distance between the marginal and conditional explanations for f_α, we obtain

$$\frac{\alpha^2}{2} \leq \frac{2}{(1+\delta^2)^2}\Big(\alpha^2 + 2\delta^2(1+\frac{\alpha^2}{4})\Big) = \|\bar{\varphi}^{ME}(X;f_\alpha) - \bar{\varphi}^{CE}(X;f_\alpha)\|^2_{L^2(\mathbb{P})} \tag{2.13}$$

assuming $\delta < 1$. Thus, if the dependencies are strong, the expression suggests that the marginal explanations approximate well the conditional explanations of the response only when α is small, in which case $f_\alpha \approx f_0$ pointwise. Moreover, whenever $|\alpha| \geq \frac{\delta^{-\gamma}}{2}$, the lower bound in (2.13) satisfies $\frac{\alpha^2}{2} \geq \frac{\delta^{-2\gamma}}{8} \to \infty$ as $\delta \to 0^+$.

Now, by taking dependencies into account, we form the two feature groups $(X_1, X_2), X_3$, which corresponds to the partition $\mathcal{P} = \{\{1,2\},\{3\}\}$ of $N = \{1,2,3\}$. In this case, the marginal and conditional quotient games coincide, that is, $v^{ME,\mathcal{P}}(S;x,X,f) = v^{CE,\mathcal{P}}(S;x,X,f)$, $S \subseteq \{1,2\}$. As a consequence, the Shapley value for these quotient games also coincide. In particular, we have

$$\begin{aligned}
&\bar{\varphi}_1^{ME,\mathcal{P}}(X;f_\alpha) = \bar{\varphi}_1^{CE,\mathcal{P}}(X;f_\alpha) = X_1 + X_2 + \alpha(\epsilon_1 - \epsilon_2) = \varphi_1^{CE,\mathcal{P}}(X;Y) + O(\delta^{1-\gamma}),\\
&\bar{\varphi}_2^{ME,\mathcal{P}}(X;f_\alpha) = \bar{\varphi}_2^{CE,\mathcal{P}}(X;f_\alpha) = X_3.
\end{aligned}$$

Moreover, we have the following stability relationship:

$$\|\bar{\varphi}^{CE,\mathcal{P}}(X;f_\alpha) - \varphi^{CE,\mathcal{P}}(X;Y)\|_{L^2(\mathbb{P})} = \|\bar{\varphi}^{ME,\mathcal{P}}(X;f_\alpha) - \varphi^{CE,\mathcal{P}}(X;Y)\|_{L^2(\mathbb{P})} = \|f_\alpha(X) - Y\|_{L^2(\mathbb{P})}. \tag{2.14}$$

As a consequence, for models whose predictions are close to the observations of the response variable, we can expect their marginal and conditional explanations to be similarly close to those of the response.

Figures 1e-1f shows the quotient marginal explanations (which also coincide with quotient conditional ones) for models $f_\alpha, \alpha \in [-3,3]$, respectively, compared to the observations of Y.

3 Group explainers with coalition structures

In our work, the game v^{CE} is referred to as conditional and v^{ME} as marginal; see (2.4) for definitions. If the predictors in X are independent, the two games coincide. Under 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 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 analysis carried out in this article is a continuation of the work from [45]. There, we focused on single feature explainers based on linear game values and derived bounds for the marginal and conditional operators, as well as for the $L^2(\mathbb{P})$-distance between them. Naturally, the next step is to derive similar results for quotient game values and coalitional values.

In this section, we construct explainers that quantify predictor contributions to the model output by considering predictor unions. In particular, given predictors $X \in \mathbb{R}^n$ and a partition $\mathcal{P} = \{S_1, S_2, \ldots, S_m\}$ of N, our objective is to explain the contribution of each predictor X_i under a coalition structure by utilizing the partition $\mathcal{P}$, as well as the contribution of each group X_{S_j}. We will refer to such explainers as explainers with a coalition structure.

Here, we demonstrate analytically that grouping features based on the strength of dependencies has a stabilizing effect on the marginal explanations at both group and individual levels, and enables the approximation of conditional explanations by marginal ones.

3.1 Stability of explanations based on quotient games

A necessary ingredient for constructing such operators is a linear game value which allows quantifying the contribution of each feature. For simplicity, in our work, we consider linear game values in the (marginalist) form

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

with $w(S,n) \geq 0$ and where S is a proper subset of N. Such game values can be extended trivially to non-cooperative games such as marginal and conditional games using the form (3.1); extensions for all types of game values are discussed in [44]. Notice that the Shapley value (2.3) 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 Appendix A.

In the case when predictors within each union X_{S_j} share significant amount of mutual information, the change in value of one of the predictors causes a certain change in value of other predictors in the union, and thus, the predictors within the union "act in agreement" with one another. Thus, constructing explainers that explicitly incorporate the coalition structure of $\mathcal{P} = \{S_1, S_2, \ldots, S_m\}$ might be advantageous when the partition is based on dependencies. To design explainers of unions with the partition in mind, we make use of quotient games defined in (2.6).

By design, the quotient game is played by the unions; that is, the game $v^{\mathcal{P}}$ is obtained by restricting v to unions $S_j \in \mathcal{P}$ by viewing the elements of the partition $\mathcal{P}$ as players. The complexity of the quotient game value $h[M, v^{\mathcal{P}}]$ is of the order $2^{|\mathcal{P}|} \cdot O(v)$, where $O(v)$ stands for the complexity of the game evaluation for any $S \subseteq N$. We also note that if $\mathcal{P}$ contains singletons, that is, $\mathcal{P} = \bar{N} := \{\{1\},\{2\},\ldots,\{n\}\}$, then $v = v^{\bar{N}}$.

3.1.1 Conditional quotient game operators on $L^2(P_X)$

An important property of the marginal and conditional games is that of linearity with respect to models, which is also true for the corresponding quotient games. Specifically, given random features $X = (X_1, X_2, \dots, X_n)$, a partition $\mathcal{P} = \{S_1, S_2, \dots, S_m\}$ of N, and two continuous models f, g we have

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

Since the game value $h[N, v]$ in (3.1) is also linear, the linearity extends to $h[M, v^{\mathcal{P}}]$ as well,

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

on the space of continuous models.

To extend the marginal and conditional quotient 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 and the corresponding game values are well-defined maps.

For the conditional game value, an appropriate space is $L^q(P_X)$, $q \geq 1$, with the corresponding L^q-norm

$$\|f\|^q_{L^q(P_X)} = \int f^q(x) P_X(dx) = \mathbb{E}[|f(X)|^q].$$

The above norm measures the distance between models by evaluating the expected difference between predictions (to the power q). For the sake of exposition we work with $q = 2$, which ensures that $Var(f(X)) < \infty$.

Definition 3.1. *Let $X = (X_1, \dots, X_n)$ be defined on $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{P} = \{S_1, S_2, \dots, S_m\}$ be a partition of N, and $h[N, v]$ be given by* (3.1)*. The conditional game operator $\bar{\mathcal{E}}^{CE}[\cdot; h, X, \mathcal{P}] : L^2(P_X) \to L^2(\Omega, \mathcal{F}, \mathbb{P})^m$ associated with $h, X, \mathcal{P}$ is defined by*

$$\bar{\mathcal{E}}^{CE}[f; h, X, \mathcal{P}] := h[M, v^{CE,\mathcal{P}}(\cdot; X, f)], \quad f \in L^2(P_X).$$

On notation. 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 $(\Omega, \mathcal{F}, \mathbb{P})$. We also assume that a partition of $N = \{1, \dots, n\}$ has the form $\mathcal{P} = \{S_1, \dots, S_m\}$, $1 \leq m \leq n$, and set $M := \{1, \dots, m\}$. Furthermore, given the partition $\mathcal{P}$ and $A \subseteq M$, we use the notation $Q_A := \cup_{k \in A} S_k$ where we suppressed the dependence on $\mathcal{P}$; otherwise it will be explicitly used when context is needed. For any partition $\mathcal{P}$ we define the collection of quotient coalitions by $\mathcal{C}(N, \mathcal{P}) := \{S \subseteq N : S = \cup_{j \in A} S_j, A \subseteq M\}$.

We now provide the bounds and list other properties for a game value of the quotient conditional game.

Proposition 3.1 (conditional bounds). *Let h, X, $\mathcal{P}$ be as in Definition 3.1.*

(i) *The map $f \in L^2(P_X) \mapsto \{v^{CE,\mathcal{P}}(A; X, f)\}_{A \subseteq M} \in (L^2(\mathbb{P}))^{2^m}$ is a well-defined, bounded linear operator satisfying*

$$\|v^{CE,\mathcal{P}}(A; X, f)\|_{L^2(\mathbb{P})} \leq \|f\|_{L^2(P_X)}, \quad A \subseteq M.$$

(ii) *$\bar{\mathcal{E}}^{CE}[\cdot; h, X, \mathcal{P}] = (\bar{\mathcal{E}}^{CE}_1, \dots, \bar{\mathcal{E}}^{CE}_m)$ is a well-defined, bounded linear operator satisfying*

$$\|\bar{\mathcal{E}}^{CE}_j[f; h, X, \mathcal{P}]\|_{L^2(\mathbb{P})} \leq \Big(\sum_{A \subseteq M \setminus \{j\}} w(A, m) \Big) \|f\|_{L^2(P_X)}, \quad f \in L^2(P_X), \quad j \in M. \tag{3.2}$$

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

(iv) *If h satisfies the efficiency property (EP), the Lipschitz inequality from (ii) can be improved as*

$$\sum_{j=1}^{m} \|\bar{\mathcal{E}}^{CE}_j[f; h, X, \mathcal{P}]\|^2_{L^2(\mathbb{P})} \leq \|f - f_0\|^2_{L^2(P_X)} \leq \|f\|^2_{L^2(P_X)}, \quad f \in L^2(P_X),\ f_0 = \mathbb{E}[f(X)]. \tag{3.3}$$

Proof. See Appendix C.1. □

Remark 3.1. Proposition 3.1(ii) suggests that if predictions of two models (on average) are close to each other, than their quotient conditional explanations (on average) will be close. Here, the Lipschitz bound in (3.2), which depends on the partition $\mathcal{P}$, determines the relative scale between explanation differences and the differences of associated models. In the case when h is efficient, (3.3) states that the differences between the (vectors of) explanations are bounded by the differences in predictions of the two models (that is, the Lipschitz bound equals 1). As we will see, this is not always the case for the marginal operator when dependencies are present.

3.1.2 Marginal quotient game operators on $L^2(\tilde{P}_X)$.

We next take a similar approach in constructing an operator based on the marginal quotient game. Let $\mathcal{P} = \{S_1, \dots, S_m\}$ be a partition of N. To choose an appropriate space of models, note that for any bounded $f \in \mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$ we have for any $A \subseteq M = \{1, \dots, m\}$

$$\mathbb{E}\big[v^{ME,\mathcal{P}}(A; X; f)\big] = \int f(x_{Q_A}, x_{-Q_A})[P_{X_{Q_A}} \otimes P_{X_{-Q_A}}](dx_{Q_A}, dx_{-Q_A}).$$

If the predictor groups $X_{S_1}, \dots, X_{S_m}$ are not independent, the product measures $P_{X_{Q_A}} \otimes P_{X_{-Q_A}}$, $A \subseteq M$, will in general differ from P_X. Hence, since the marginal explanations based on the game value (3.1) are linear combinations of $v^{ME,\mathcal{P}}(A; X, f)$, $A \subseteq M$, natural domains for the quotient marginal operator are the spaces $L^q(\tilde{P}_{X,\mathcal{P}})$, $q \geq 1$, with the corresponding co-domains being $L^q(\mathbb{P})$, where

$$\tilde{P}_{X,\mathcal{P}} := \frac{1}{2^m} \sum_{A \subseteq M} P_{X_{Q_A}} \otimes P_{X_{-Q_A}} \tag{3.4}$$

and the corresponding L^q-norm is given by

$$\|f\|^q_{L^q(\tilde{P}_{X,\mathcal{P}})} := \frac{1}{2^m} \sum_{A \subseteq M} \int f(x_{Q_A}, x_{-Q_A})[P_{X_{Q_A}} \otimes P_{X_{-Q_A}}](dx_{Q_A}, dx_{-Q_A}), \tag{3.5}$$

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, for simplicity, we develop the L^2-theory for the marginal quotient explanations.

Remark 3.2. If the predictor groups $X_{S_1}, \dots, X_{S_m}$ are independent (see §2.1), then $\tilde{P}_{X,\mathcal{P}} = P_X$. In that case, $v^{ME,\mathcal{P}} = v^{CE,\mathcal{P}}$ and hence $h[N, v^{ME,\mathcal{P}}]$ is a well-defined linear operator on $L^2(P_X)$ which $\mathbb{P}$-a.s. equals to $h[N, v^{CE,\mathcal{P}}]$. Group independence, however, is a very stringent requirement. Thus, when there is no knowledge on dependencies, the most suitable probability measure for the marginal quotient game is $\tilde{P}_{X,\mathcal{P}}$. We also note that in the case of singletons, when $\mathcal{P} = \bar{N}$, we have $\tilde{P}_{X,\bar{N}} = \tilde{P}_X$.

Definition 3.2. *Let h, X, $\mathcal{P}$ be as in Definition 3.1. The marginal game operator $\bar{\mathcal{E}}^{ME} : L^2(\tilde{P}_{X,\mathcal{P}}) \to L^2(\mathbb{P})^m$ associated with $h, X, \mathcal{P}$ is defined by*

$$\bar{\mathcal{E}}^{ME}[f; h, X, \mathcal{P}] := h[M, v^{ME,\mathcal{P}}(\cdot; X, f)], \quad f \in L^2(\tilde{P}_{X,\mathcal{P}}). \tag{3.6}$$

Proposition 3.2 (marginal bounds I). *Let h, X, $\mathcal{P}$ be as in Definition 3.1.*

(i) *The map $f \in L^2(\tilde{P}_{X,\mathcal{P}}) \mapsto \{v^{ME,\mathcal{P}}(A; X, f)\}_{A \subseteq M} \in (L^2(\mathbb{P}))^{2^m}$ is a well-defined, bounded linear operator satisfying*

$$\|v^{ME,\mathcal{P}}(A; X, f)\|_{L^2(\mathbb{P})} \leq \|f\|_{L^2(P_{X_{Q_A}} \otimes P_{X_{-Q_A}})}, \quad A \subseteq M. \tag{3.7}$$

(ii) *$\bar{\mathcal{E}}^{ME}[\cdot; h, X, \mathcal{P}] = (\bar{\mathcal{E}}^{ME}_1, \dots, \bar{\mathcal{E}}^{ME}_m)$ is a well-defined, bounded linear operator on $L^2(\tilde{P}_{X,\mathcal{P}})$ satisfying*

$$\|\bar{\mathcal{E}}^{ME}_j[f; h, X, \mathcal{P}]\|_{L^2(\mathbb{P})} \leq 2^{\frac{m+1}{2}} \Big(\sum_{A \subseteq M \setminus \{j\}} w^2(A, m) \Big)^{\frac{1}{2}} \|f\|_{L^2(\tilde{P}_{X,\mathcal{P}})}, \quad f \in L^2(\tilde{P}_{X,\mathcal{P}}), \quad j \in M. \tag{3.8}$$

(iii) $\big\{f \in L^2(\tilde{P}_{X,\mathcal{P}}) : f = const\ \tilde{P}_{X,\mathcal{P}}\text{-}a.s.\big\} \subseteq \mathrm{Ker}(\bar{\mathcal{E}}^{ME}[\cdot; h, X, \mathcal{P}])$.

Proof. See Appendix C.2. □

The bound in Theorem 3.2(*ii*) implies that the marginal quotient explanations are continuous, or stable, in $L^2(\tilde{P}_{X,\mathcal{P}})$. However, as we will see in §3.1.3, two models that are close in $L^2(P_X)$ may yield marginal quotient explanations that are far apart in that space when dependencies exist among predictor groups. In other words, the map $f \mapsto \bar{\mathcal{E}}^{ME}[f; X, h, \mathcal{P}]$ may be unbounded on domains equipped with the joint probability measure P_X (when it is well-defined). That instability can be quantified, as we will also see, by bounding the L^2-distance between conditional and marginal quotient games where the bound considers the strength of the dependencies across groups.

3.1.3 Marginal quotient game operator on $L^2(P_X)$

The objective of this section is to investigate the marginal quotient explanations in the space where the distance between models is measured as the distance between predictions, that is, the space equipped the with $L^2(P_X)$-norm. This will help us understand how grouping dependent features impacts the stability of marginal quotient explanations and when they serve as good approximations to conditional ones.

To further understand why the $L^2(P_X)$-norm is in general less suitable for the domain of the marginal operator, consider the following example.

Example 3.1. Let $X = (X_1, X_2, X_3)$, $N = \{1, 2, 3\}$, and suppose P_X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^3$ with the density $p(x_1, x_2, x_3) = c_*(x_1 x_2^4 + x_1^4 x_2)\mathbb{1}_{[0,1]^3}(x_1, x_2, x_3)$, where c_* is the normalization constant. First, let us consider single feature explanations. By construction we have $\tilde{P}_X \ll P_X$ and $P_X \ll \tilde{P}_X$. Thus, the P_X-equivalence and $\tilde{P}_X$-equivalence classes coincide, allowing to equip the $L^2(\tilde{P}_X)$ space with the $L^2(P_X)$-norm.

Next, set $g^{(t)}(x_1, x_2, x_3) = \mathbb{1}_{[0,t]^2}(x_1, x_2)$, $t \in (0, 1)$. Take any $f_1 \in L^2(\tilde{P}_X)$ and let $f_2^{(t)} = f_1 + g^{(t)}$. Then, by (2.3), for $i \in \{1, 2\}$ we obtain

$$\|\varphi_i[N, v^{ME,\mathcal{P}}(\cdot; X, h, f_1 - f_2^{(t)})]\|^2_{L^2(\mathbb{P})} \geq \frac{1}{2} \cdot (t^{-1} - 1)\|f_1 - f_2^{(t)}\|_{L^2(P_X)} \to \infty \quad \text{as } t \to 0^+.$$

This illustrates that the marginal Shapley values can be unbounded in $L^2(P_X)$; see [45] for other examples.

We next define $\mathcal{P} = \{\{1, 2\}, \{3\}\}$. In light of independence of (X_1, X_2) from X_3 we have $\tilde{P}_{X,\mathcal{P}_*} = P_X$. Thus, the quotient games $v^{ME,\mathcal{P}}$ and $v^{CE,\mathcal{P}}$ coincide. Hence for any $f_1, f_2 \in L^2(P_X)$

$$\|\varphi[N, v^{CE,\mathcal{P}}(\cdot; X, f_1 - f_2)]\|^2_{L^2(\mathbb{P})} = \|\varphi[N, v^{ME,\mathcal{P}}(\cdot; X, f_1 - f_2)]\|^2_{L^2(\mathbb{P})} \leq \|f_1 - f_2\|_{L^2(P_X)}.$$

Therefore grouping by dependencies allows us to bound the quotient marginal Shapley values in $L^2(P_X)$-norm.

If one attempts to equip the space $L^2(\tilde{P}_{X,\mathcal{P}})$ with the $L^2(P_X)$-norm, then the marginal quotient game operator may not always be well-defined. To this end, we define the following function space.

Definition 3.3. *Let $X, \mathcal{P}$ be as in Definition 3.1. Define the space associated with $(X, \mathcal{P})$ as follows:*

$$H_{X,\mathcal{P}} := \left(\left\{[f] : [f] = \left\{\tilde{f} : \tilde{f} = f\ P_X\text{-}a.s.\ and \int |\tilde{f}(x)|^2 \tilde{P}_{X,\mathcal{P}}(dx) < \infty\right\}\right\}, \|\cdot\|_{L^2(P_X)}\right) \hookrightarrow L^2(P_X). \quad (3.9)$$

When $\mathcal{P} = \bar{N}$, we set $H_X := H_{X,\bar{N}}$.

It is crucial to point out that $P_X \ll \tilde{P}_{X,\mathcal{P}}$ and hence if $f_1 = f_2$ $\tilde{P}_{X,\mathcal{P}}$-a.s., then $f_1 = f_2$ P_X-almost surely. For this reason, either $H_{X,\mathcal{P}}$ contains exactly the same elements as $L^2(\tilde{P}_{X,\mathcal{P}})$ or some elements of $L^2(\tilde{P}_{X,\mathcal{P}})$ are placed in the same equivalence class of $H_{X,\mathcal{P}}$. As the next lemma shows, the former happens if $\tilde{P}_{X,\mathcal{P}} \ll P_X$, which allows for the marginal quotient explanations to be well-defined on $H_{X,\mathcal{P}}$.

Lemma 3.1 (well-posedness)**.** *Let $X, \mathcal{P}$ be as in Definition 3.1. Suppose $\tilde{P}_{X,\mathcal{P}} \ll P_X$.*

(i) $H_{X,\mathcal{P}} \cong (L^2(\tilde{P}_{X,\mathcal{P}}), \|\cdot\|_{L^2(P_X)})$.

(ii) The map $f \in H_{X,\mathcal{P}} \mapsto \{v^{\scriptscriptstyle ME,\mathcal{P}}(A; X, f)\}_{A\subseteq M} \in (L^2(\mathbb{P}))^{2^m}$ is well-defined.

(iii) $(\bar{\mathcal{E}}^{\scriptscriptstyle ME}[\cdot; h, X, \mathcal{P}], H_{X,\mathcal{P}})$ acting via the formula (3.6) with h as in (3.1) is well-defined.

Proof. See Appendix C.3. □

The above lemma states that if the density of $\tilde{P}_{X,\mathcal{P}}$ with respect to P_X exists, then the marginal quotient game value as an operator on $H_{X,\mathcal{P}}$ is well-defined. It can be shown that the absolute continuity is a necessary and sufficient condition for well-posedness of the marginal quotient game on $H_{X,\mathcal{P}}$. However, for the sake of exposition, we omit the proof of this fact; see [45] where we prove this in the case of singletons $\mathcal{P} = \bar{N}$.

When $\tilde{P}_{X,\mathcal{P}} \ll P_X$ the Radon-Nikodym derivative of $\tilde{P}_{X,\mathcal{P}}$ with respect to P_X exists and encodes information about feature dependencies. The following lemma, which will be useful for our analysis, provides a representation of the Radon-Nikodym derivative and the spaces $L^2(\tilde{P}_{X,\mathcal{P}})$ and $H_{X,\mathcal{P}}$.

Lemma 3.2. *Suppose $\tilde{P}_{X,\mathcal{P}} \ll P_X$. Let $r^{(\mathcal{P})} := \frac{d\tilde{P}_{X,\mathcal{P}}}{dP_X}$. Then:*

(i) $L^2(\tilde{P}_{X,\mathcal{P}})$ can be identified with the weighted L^2-space $L^2_{r^{(\mathcal{P})}}(P_X)$ where

$$r^{(\mathcal{P})} = \tfrac{1}{2^m} \sum_{A\subseteq M} r_{Q_A} \geq \tfrac{1}{2^{m-1}}, \quad \textit{where} \;\; 0 \leq r_{Q_A} := \frac{dP_{X_{Q_A}} \otimes P_{X_{-Q_A}}}{dP_X} \in L^1(P_X), \;\; \|r_{Q_A}\|_{L^1(P_X)} = 1. \tag{3.10}$$

(ii) $H_{X,\mathcal{P}} = L^2(P_X)$ if and ony if $r^{(\mathcal{P})} \in L^\infty(P_X)$, that is, $r_{Q_A} \in L^\infty(P_X)$, $A \subseteq M$.

Proof. See Appendix C.4. □

In our companion paper [45] where we investigate the stability of individual (non-quotient) marginal explanations we do an in-depth analysis of ill-posedness and boundedness of the marginal game on the space H_X. In this article, however, to make the exposition simpler and to avoid repetitive conclusions, we consider only cases where the marginal quotient game is well-defined, bounded, and can be placed in the same space together with the conditional quotient game in order to get approximation bounds. To this end, we list the following useful assumptions which we will be employing throughout the article:

(AC) $\tilde{P}_X \ll P_X$, that is, $P_{X_S} \otimes P_{X_{-S}} \ll P_X$, for every $S \subseteq N$.

(PB) Given partition $\mathcal{P}$ of N, we assume $r^{(\mathcal{P})} \in L^\infty$, that is, $r_{Q_A} \in L^\infty(P_X)$, $A \subseteq M$.

The above assumptions are not too stringent, which allow us to demonstrate the stabilizing mechanisms of grouping in a simple manner without making the material excessively complicated. In what follows, for the sake of the exposition, we suppress the dependence on X in the notation of the operators $\bar{\mathcal{E}}^{\scriptscriptstyle CE}, \bar{\mathcal{E}}^{\scriptscriptstyle ME}$.

The absolute continuity condition (AC) together with (PB) allows the Radon-Nikodym derivative $r^{(\mathcal{P})}$ to control the strength of dependencies among the predictors. To see this, we establish the bound on the distance between the probability measures $\tilde{P}_{X,\mathcal{P}}$ and P_X that relies on $r^{(\mathcal{P})}$.

Lemma 3.3. *Let $\mathcal{P}$, $r^{(\mathcal{P})}$, and r_{Q_A} be as in Lemma 3.2. Suppose (AC) and (PB) hold, and $\mathbb{E}[|X|] < \infty$.*

$$\begin{aligned} \left| \int |x| \cdot (r^{(\mathcal{P})}(x) - 1)\, P_X(dx) \right| &\leq W_1(\tilde{P}_{X,\mathcal{P}}, P_X) \\ &\leq \int |x| \cdot |r^{(\mathcal{P})}(x) - 1|\, P_X(dx) \leq \frac{\mathbb{E}[|X|]}{2^m} \sum_{A\subseteq M} \|r_{Q_A}(x) - 1\|_{L^\infty(\mathbb{P})} < \infty. \end{aligned} \tag{3.11}$$

Proof. Follows from (PB), Lemma 3.2, Lemma D.2, and the triangle inequality. □

Lemma 3.11 implies that if $r^{(\mathcal{P})}=1$, then $r_{Q_A}=1$, $A\subseteq M$ and the two measures coincide. When $r^{(\mathcal{P})}$ deviates from 1, the dependencies start to impact the distance. As a consequence, the quotient marginal and conditional explanations deviate from each another, as discussed in the approximation estimate below.

Proposition 3.3 (approximation). *Let X, $\mathcal{P}$, h be as in Definition 3.1 and $r^{(\mathcal{P})}$, r_{Q_A} as in Lemma 3.2. Suppose (AC) and (PB) hold.*

(i) For every $f\in H_{X,\mathcal{P}}=L^2(P_X)$ we have

$$\|v^{CE,\mathcal{P}}(A;X,f)-v^{ME,\mathcal{P}}(A;X,f)\|_{L^2(\mathbb{P})}\le\|r_{Q_A}-1\|_{L^\infty(P_X)}\|f\|_{L^2(P_X)},\quad A\subseteq M.$$

(ii) In the presence of dependencies between predictor groups, we have for $f\in L^2(P_X)=H_{X,\mathcal{P}}$

$$\bar{\mathcal{E}}_j^{CE}[f;h,\mathcal{P}]=\bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}]+\mathcal{I}_j(f;\{r_{Q_A}\}_{A\subseteq M},h)\quad in\quad L^2(\mathbb{P}),\tag{3.12}$$

where the error term $\mathcal{I}_j$ satisfies the following approximation bound

$$\|\mathcal{I}_j(f;\{r_{Q_A}\}_{A\subseteq M},h)\|_{L^2(\mathbb{P})}\le2\Big(\sum_{A\subseteq M\backslash\{j\}}w(A,m)\Big)\cdot\Big(\max_{A\subseteq M}\|r_{Q_A}-1\|_{L^\infty(P_X)}\Big)\cdot\|f\|_{L^2(P_X)}.\tag{3.13}$$

Proof. See Appendix C.5. □

Proposition 3.3 implies that if the predictor groups $X_{S_1},X_{S_2},\dots,X_{S_m}$ are independent, then $r_{Q_A}=1$, $A\subseteq M$, and hence the marginal and conditional quotient games coincide, that is, $v^{CE,\mathcal{P}}=v^{ME,\mathcal{P}}$. Consequently, we have $\bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}]=\bar{\mathcal{E}}_j^{CE}[f;h,\mathcal{P}]$, $j\in M$.

The approximation bounds in Proposition 3.3 imply that the marginal quotient game operator is continuous in $L^2(P_X)$ and satisfies the following growth bounds.

Proposition 3.4 (marginal bounds II). *Let X, $\mathcal{P}$, h be as in Definition 3.1 and $r^{(\mathcal{P})}$, r_{Q_A} as in Lemma 3.2. Suppose (AC) and (PB) hold.*

(i) The map $f\in H_{X,\mathcal{P}}=L^2(P_X)\mapsto\{v^{ME,\mathcal{P}}(A;X,f)\}_{A\subseteq M}\in(L^2(\mathbb{P}))^{2^m}$ is a well-defined, bounded linear operator satisfying

$$\|v^{ME,\mathcal{P}}(A;X,f)\|_{L^2(\mathbb{P})}\le\big(1+\|1-r_{Q_A}\|_{L^\infty(P_X)}\big)\|f\|_{L^2(P_X)},\quad A\subseteq M.$$

(ii) The map $f\in H_{X,\mathcal{P}}=L^2(P_X)\mapsto\bar{\mathcal{E}}^{ME}[f;h,\mathcal{P}]\in(L^2(\mathbb{P}))^m$ is a well-defined, bounded linear operator satisfying for $j\in M$

$$\|\bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})}\le\Big(1+2\cdot\max_{A\subseteq M}\|r_{Q_A}-1\|_{L^\infty(P_X)}\Big)\Big(\sum_{A\subseteq M\backslash\{j\}}w(A,m)\Big)\|f\|_{L^2(P_X)}.\tag{3.14}$$

(iii) If h satisfies the efficiency property (EP), the Lipschitz inequality from (ii) can be improved as

$$\|\bar{\mathcal{E}}^{ME}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})^m}\le\Big(1+2\cdot\sqrt{m}\cdot\max_{A\subseteq M}\|r_{Q_A}-1\|_{L^\infty(P_X)}\Big)\|f\|_{L^2(P_X)}.\tag{3.15}$$

Proof. See Appendix C.6. □

Grouping as a stabilizing mechanism. The bounds in Proposition 3.4 allow us to demonstrate that grouping serves as the stabilizing mechanism when using marginal quotient explanations. Suppose $\mathcal{P}$ is a partition with $|\mathcal{P}|<n$. Suppose (AC) holds and (PB) holds for both $\mathcal{P}$ and the partition of singletons

$\bar{N}$, in which case $H_{X,\mathcal{P}} = H_{X,\bar{N}} = L^2(P_X)$. Then the bounds (3.15) for the vector of marginal quotient explanations for $\mathcal{P}$ and $\bar{N}$ satisfy the relationship

$$\Big(1 + 2 \cdot \sqrt{|\mathcal{P}|} \cdot \max_{S \subseteq \mathcal{C}(N,\mathcal{P})} \|r_S - 1\|_{L^\infty(P_X)}\Big) < \Big(1 + 2 \cdot \sqrt{n} \cdot \max_{S \subseteq N} \|r_S - 1\|_{L^\infty(P_X)}\Big) \tag{3.16}$$

where by definition $\mathcal{C}(N,\mathcal{P}) = \{S \subseteq N : S = \cup_{j \in A} S_j, A \subseteq M\}$. Additionally, if (PB) holds only for $\mathcal{P}$, for some $S \notin \mathcal{C}(N,\mathcal{P})$ one might have that $r_S \notin L^\infty(P_X)$, or even $v^{ME}(S; f, X)$, might be unbounded on H_X. In the latter case, no matter how close the predictions of the two models f_1, f_2 are, the distance between their marginal game values $v^{ME}(S; f_1, X)$ and $v^{ME}(S; f_2, X)$ may be arbitrarily large.

More generally, suppose we have a nested sequence of partitions $\{N\} \prec \mathcal{P}_1 \prec \mathcal{P}_2 \prec \cdots \prec \{\bar{N}\}$ for which (PB) holds. These type of partitions often arise in hierarchical clustering of variables based on mutual information-based metrics (see §4); in this case, the strength of dependencies between groups is increasing with the size of the partition. Whatever the case, the bounds for the vector of explanations increase in the same order because $\mathcal{C}(N,\{N\}) \subset \mathcal{C}(N,\mathcal{P}_1) \subset \mathcal{C}(N,\mathcal{P}_2) \cdots \subset \mathcal{C}(N,\bar{N})$. While the analysis of bounds clearly demonstrates theoretical improvement in stability, our numerical experiments demonstrate that the dissimilarity between model explanations decreases significantly when groups are formed based on dependencies (see §5).

Implications on feature importance. Propositions 3.1(ii) and 3.3(i) have a direct implication on any processes that make use of local feature group attributions to make data-informed decisions. Suppose $f_1, f_2 \in L^2(P_X)$ are two distinct models trained on the same dataset that generate similar predictions and the partition $\mathcal{P}$ indicates the predictor groups based on dependencies. Then the aforementioned results state that the marginal quotient explanations $\bar{\mathcal{E}}^{ME}[f_1; h, \mathcal{P}]$ and $\bar{\mathcal{E}}^{ME}[f_2; h, \mathcal{P}]$ will be close in an $L^2(\mathbb{P})$-sense. To understand the importance of this, consider any case where models are frequently retrained and their outputs impact every-day people, such as credit risk models from financial institutions. By adhering to predictor groups that are based on dependencies, the marginal quotient explanations will remain consistent across distinct models when the underlying data have not drifted from their original empirical distribution. In turn, this provides high fidelity information to companies and consumers alike.

Using the quotient game approach, there are certain considerations one must take into account:

(a) if the partition is changed the game values have to be recomputed;

(b) knowledge of quotient game values cannot provide information on single feature explanations, which are expensive to compute;

(c) even if single feature explanations are known, the trivial and quotient game explanations in general are not equal; this case causes loss of continuity of marginal, trivial group explanations with respect to models in $L^2(P_X)$ when dependencies are present.

The aforementioned difficulties can be overcome when explainers are constructed with the help of coalitional values that utilize the partition structure for computing single feature explanations. We discuss such explainers in the next section.

3.2 Group explanations based on the coalitional value operator on $L^2(P_X)$

A more advanced way to design explainers that consider partitions of the predictor index set N is to employ cooperative game theory with coalition structure, in which the objective is to compute the payoffs of players in a game where players form unions and act in agreement within the union.

Games with coalitions were introduced by [7] and later many more researchers contributed to the development of this subject. Some of the notable works are [46], [47], [62], [17], [4], [5], [3], [11], [64]. See also the work by Lorenzo-Freire [39] containing a detailed exposition on games with coalitions.

Definition 3.4. *Let $N \subset \mathbb{N}$ and $\mathcal{P} = \{S_1, S_2, \dots, S_m\}$ be a partition of N. A coalitional value g is a map that assigns to every game with a coalition structure $(N, v, \mathcal{P})$ a vector*

$$g[N, v, \mathcal{P}] = \{g_i[N, v, \mathcal{P}]\}_{i \in N}$$

where $g_i[N, v, \mathcal{P}]$ denotes the payoff for the player $i \in N$.

Properties of game values such as linearity (LP), efficiency (EP) etc. (see Appendix A) extend to coalitional game values in an obvious way. In what follows, for the sake of exposition, we consider coalitional values in the following form.

Definition 3.5. *Given the weight functions $w^{(1)}(S, |N|) \geq 0$ and $w^{(2)}(S, |N|) \geq 0$ defined for every $N \subset \mathbb{N}$ and $S \subseteq N$, we set $w := (w^{(1)}, w^{(2)})$ and define a coalition value by*

$$g_i^w[N, v, \mathcal{P}] = \sum_{A \subseteq M \setminus \{j\}} \sum_{T \subseteq S_j \setminus \{i\}} w^{(1)}(|A|, |M|) w^{(2)}(|T|, |S_j|) \big(v(Q_A \cup T \cup \{i\}) - v(Q_A \cup T)\big), \quad i \in S_j, \tag{3.17}$$

where $\mathcal{P} = \{S_1, \dots, S_m\}$ and $M = \{1, \dots, m\}$, and require that $w^{(1)}(0,1), w^{(2)}(0,1) > 0$.

Some notable coalitional values in the above form are the Owen value and the Banzhaf-Owen value respectively defined by the weights

$$w_{Ow}^{(1)} = \frac{a!(m-a-1)!}{m!}, \; w_{Ow}^{(2)} = \frac{t!(s_j - t - 1)!}{s_j!}, \; \text{and } w_{BzOw}^{(1)} = \frac{1}{2^{m-1}}, \; w_{BzOw}^{(2)} = \frac{1}{2^{s_j - 1}} \tag{3.18}$$

where $t = |T|$, $s_j = |S_j|$ and $a = |A|$. The difference between the two values is that the Owen value satisfies the efficiency property, while the Banzhaf-Owen value satisfies the total power property. In addition, the Owen value for partitions consisting of singletons is the Shapley value (2.3), while for such partitions the Banzhaf-Owen value is the Banzhaf value [8]. These properties can be verified directly.

The coalitional value in the form (3.17) naturally induces two linear game values associated with coefficients $w^{(k)}, k \in \{1, 2\}$, which are unique up to a scaling constant. These are formally defined below.

Definition 3.6. *Let a coalitional value g^w and the weights $w = (w^{(1)}, \, w^{(2)})$ be as in (3.17). We set $\alpha_* := w^{(1)}(0,1) \cdot w^{(2)}(0,1)$, and define the linear game values $h_*^{(1)}$ and $h_*^{(2)}$ induced by g^w as follows:*

$$h_{*,i}^{(k)}[N, v] = \frac{1}{w^{(k)}(0,1)} \sum_{S \subseteq N \setminus \{i\}} w^{(k)}(|S|, |N|) \big(v(S \cup \{i\}) - v(S)\big), \quad i \in N, \, N \subset \mathbb{N}, \, k \in \{1, 2\}, \tag{3.19}$$

where the renormalization factor $1/w^{(k)}(0,1)$ in (3.19) ensures that $h_{*,i}^{(k)}[\{i\}, v] = v(i)$.

We note that for the Owen and Banzhaf-Owen values we have $h_*^{(1)} = h_*^{(2)} = \varphi$ and $h_*^{(1)} = h_*^{(2)} = Bz$, respectively, and that the normalizing constant $\alpha_* = 1$ in both cases.

Definition 3.7. *Let $X, \mathcal{P}$ be as in Definition 3.1. Let g^w be as in Definition 3.5. A trivial explainer of $f \in \mathcal{C}_{\mathcal{B}(\mathbb{R}^n)}$ based on g^w is defined by*

$$g_{S_j}^w(f, v^{ME}, \mathcal{P}; X) := \sum_{i \in S_j} g_i^w(f, v^{ME}, \mathcal{P}; X), \quad S_j \in \mathcal{P}, \quad j \in M,$$

where $g_i^w(f, v^{ME}, \mathcal{P}; X) := g[N, v^{ME}(\cdot; X, f), \mathcal{P}]$.

Trivial group explanations for linear game values in general may differ between the marginal and conditional games (even if groups are independent), which can break the continuity of the corresponding marginal group explanations with respect to models in $L^2(P_X)$. However, if $h_*^{(2)}$ is an efficient game value, then trivial group explainers based on coalitional values of the form (3.17) turn out to be quotient game values, as shown below. The implication of this result is that by summing marginal coalitional values over the elements of a predictor group, given that the partition has been formed based on dependencies, will generate the marginal quotient game value for that group and will be equal to the corresponding conditional value. In other words, one can obtain stable marginal group explanations by summing marginal coalitional values under the aforementioned conditions.

Proposition 3.5. *Let X, $\mathcal{P}$ be as in Definition 3.1 and $r^{(\mathcal{P})}$, r_{Q_A} as in Lemma 3.2. Suppose (AC) and (PB) hold. Let g^w and the weights $w = (w^{(1)}, w^{(2)})$ be as in* (3.17)*. Let α_*, $h_*^{(1)}$, $h_*^{(2)}$ be induced by g^w as in Definition 3.6. Suppose $h_*^{(2)}$ satisfies the efficiency property (EP). Then:*

(i) For any $(N, v, \mathcal{P})$ we have

$$\sum_{i \in S_j} g_i[N, v, \mathcal{P}] = \alpha_* h_{*,j}^{(1)}[M, v^{\mathcal{P}}] = g[M, v^{\mathcal{P}}, \bar{M}]. \tag{3.20}$$

(ii) For every $f \in L^2(P_X) = H_{X,\mathcal{P}}$ we have the following approximation bounds

$$g_{S_j}^w(f, v^{CE}; X) = g_{S_j}(f, v^{ME}; X) + \mathcal{I}_j(f; \{r_{Q_A}\}_{A \subseteq M}, h_*^{(1)}) \quad in \quad L^2(\mathbb{P}), \tag{3.21}$$

where the error term $\mathcal{I}_j$ satisfies the following approximation bound

$$\|\mathcal{I}_j(f; \{r_{Q_A}\}_{A \subseteq M}, h)\|_{L^2(\mathbb{P})} \leq \alpha_* \cdot 2 \Big(\sum_{A \subseteq M \setminus \{j\}} w^{(1)}(A, m) \Big) \cdot \Big(\max_{A \subseteq M} \|r_{Q_A} - 1\|_{L^\infty(P_X)} \Big) \cdot \|f\|_{L^2(P_X)}. \tag{3.22}$$

(iii) The maps $f \in H_{X,\mathcal{P}} = L^2(P_X) \mapsto g_{S_j}^w(v, \mathcal{P}, f; X) \in L^2(\mathbb{P})$, for $v \in \{v^{CE}, v^{ME}\}$, are well-defined, bounded operators satisfying for each $j \in M$

$$\begin{aligned} &\|g_{S_j}^w(f, v^{CE}, \mathcal{P}; X)\|_{L^2(\mathbb{P})} \leq \alpha_* \Big(\sum_{A \subseteq M \setminus \{j\}} w^{(1)}(A, m) \Big) \|f\|_{L^2(P_X)} \\ &\textit{and} \\ &\|g_{S_j}^w(f, v^{ME}, \mathcal{P}; X)\|_{L^2(\mathbb{P})} \\ &\qquad \leq \alpha_* \Big(1 + 2 \cdot \max_{A \subseteq M} \|r_{Q_A} - 1\|_{L^\infty(P_X)} \Big) \Big(\sum_{A \subseteq M \setminus \{j\}} w^{(1)}(A, m) \Big) \|f\|_{L^2(P_X)}. \end{aligned} \tag{3.23}$$

(iv) If $h_*^{(1)}$ satisfies the efficiency property (EP), the Lipschitz inequality from $(3.23)_1$ *can be improved as*

$$\sum_{j \in M} \|g_{S_j}(f, v^{CE}, \mathcal{P}; X)\|_{L^2(\mathbb{P})}^2 \leq \alpha_*^2 \|f - f_0\|_{L^2(P_X)}^2 \leq \alpha_*^2 \|f\|_{L^2(P_X)}^2, \quad f \in L^2(P_X),\ f_0 = \mathbb{E}[f(X)]. \tag{3.24}$$

Proof. The property (i) follows directly from Lemma B.1(i). The property (ii) follows from (i) and Proposition 3.3, while (iii) and (iv) follow from (i), Proposition 3.1, and Proposition 3.4. □

Remark 3.3. Given a scaled efficient game value $h^{(2)}$, Proposition 3.5(i) implies that

$$\sum_{i \in S_j} g_i[N, v, \mathcal{P}] = g_j[M, v^{\mathcal{P}}, \bar{M}], \quad S_j \in \mathcal{P}, \quad \mathcal{P} = \{S_1, S_2, \dots, S_m\} \tag{QP}$$

called the quotient game property. Thus, the two-step formulation of g, together with the efficiency of $h^{(2)}$, is equivalent to g satisfying (QP). Note that we could have imposed the condition (QP) on a coalitional value in order to obtain the subsequent stability results.

4 Information-theoretic hierarchical clustering of predictors

The first step in constructing group explainers is to identify disjoint sets $S_j \subseteq N$ that yield a partition $\mathcal{P} = \{S_1, S_2, \dots, S_r\}$ of predictor indices, so that, given predictors $X \in \mathbb{R}^n$, $X_{S_1}, X_{S_2}, \dots, X_{S_r}$ form (weakly) independent unions (see the definition in §2.1) where within each group the predictors share a significant amount of mutual information [15]. Such partitioning would effectively reduce the dimensionality of the problem and, consequently, lower the complexity of explanations, while also alleviating the issue of explanation instability.

Group attribution methods have previously been discussed in the context of linear or simple functional dependencies [1]. In real datasets, however, dependencies are often highly non-linear. Thus, to construct a dependency-based partition of predictors, we propose a variable hierarchical clustering technique that employs a state-of-the-art, information-theoretic measure of dependence called the Maximal Information Coefficient (MIC), that overcomes the disadvantages of traditional measures and was introduced in Reshef et al. [49, 50]. An example that demonstrates the advantage of using MIC in clustering is provided in §4.3.

4.1 Maximal information coefficient as a measure of dependence

Many notable measures of dependence have been defined in recent years: Kraskov et al. [36], Zenga [67] and Paninski [48] on the estimation of mutual information; Rényi [54] and Breiman and Friedman [9] on maximal correlation; Szekely et al. [59] and Szekely and Rizzo [60] on distance correlation; [24, 25] on the Hilbert-Schmidt independence criterion, Lopez-Paz et al. [38] on the randomized dependence coefficient; Heller et al. [29] on the Heller-Heller-Gorfine distance, Heller [30] on S^{DDP}.

Reshef et al. [50] introduced the information-theoretic measure of dependence called MIC_*, the population value of the MIC statistic, defined as a regularized form of mutual information between a pair of random variables.

Definition 4.1 (Reshef et al. [50]). *Let (X, Y) be jointly distributed random variables. The population maximal information coefficient$_*$ (MIC_*) of (X, Y) is defined by*

$$\text{MIC}_*(X,Y) = \sup_G \frac{I\big((X,Y)|_G\big)}{\log \|G\|}.$$

Here G denotes a two-dimensional grid, $\|G\|$ denotes the minimum of the number of rows of G and the number of columns of G, $I((X,Y)|_G)$ denotes the discrete mutual information of $(X,Y)|_G := (col_G(X), col_G(Y))$.

MIC_* has the following remarkable properties: (a) it returns a value in the unit interval that represents the strength of the relationship and it is 0 if and only if the variables are independent, (b) it provides a similar value when the variables are transformed via strictly monotonic functions (transitivity), and (c) it provides a similar value between pairs of variables that exhibit similar noise levels (equitability).

There are two statistics, MIC and MIC_e, that can be used to estimate MIC_*. While both are consistent estimators, MIC from Reshef et al. [49] can be computed only via an inefficient heuristic approximation, while MIC_e introduced in Reshef et al. [50] can be computed exactly and efficiently using an appropriate optimization technique that allows one to estimate MIC_* in linear time; see Definition E.1 of MIC_e and Corollary E.1 that discusses its complexity.

4.2 Dependency-based hierarchical clustering

In our work, we design a hierarchical clustering algorithm that generates a dendrogram (or partition tree) which encodes the strength of the dependencies between predictors; for details on clustering, see [28, Section 14.3]. For an example of a dendrogram, see Figure 4. To this end, we propose to use the dissimilarity measure between predictors based on MIC given by $d_{\text{MIC}_*}(X_i, X_j) = 1 - \text{MIC}_*(X_i, X_j) \in [0,1]$, estimated by the statistic $1 - \text{MIC}_e\big(\{x_i^{(\ell)}\}_{\ell=1}^M, \{x_j^{(\ell)}\}_{\ell=1}^M\big)$ based on observations $\{(x_1^{(\ell)}, \dots, x_n^{(\ell)})\}_{\ell=1}^M$.

The advantage of the MIC-based clustering algorithm is that properties of MIC are carried over to the partition tree. In particular, (a) the tree height, representing the strength of dependencies in predictors, is always ≤ 1; (b) in light of transitivity, the geometry of the tree is invariant under strictly monotone transformations; and (c) in light of equitability, the height of each subtree reflects information about the noise level among predictors corresponding to the terminal nodes of the subtree.

Given an MIC-based dendrogram of height h, the parameter $\alpha \in (0, h)$ characterizing the strength of dependencies induces a partition of predictors $\mathcal{P}_\alpha = \{S_1^\alpha, S_2^\alpha, \dots, S_{m_\alpha}^\alpha\}$ whose elements correspond to the terminal nodes of subtrees obtained by cross-sectioning the tree at height α. Under the assumption that coalescence of branches happens at distinct heights, $\alpha \mapsto \mathcal{P}_\alpha$ is a left-continuous partition map which characterizes the dendrogram and gives rise to a nested sequence of partitions starting at singletons $\{\{X_1\}, \{X_2\}, \dots, \{X_n\}\}$ and terminating at the grand coalition $\{X_1, \dots, X_n\}$. In what follows, these dependency-based partitions are used to construct group explainers based on coalitional values, which incorporate the partition into their structure. For an illustration on the use of hierarchical clustering to produce partitions and construct group explainers, see the example below.

4.3 Example of variable hierarchical clustering based on MIC$_*$

In this section we perform variable clustering using MIC$_e$ and compare it with that based on the correlation for the model:

$$\begin{array}{llll} X_0 \sim Unif(-4\pi, 4\pi), & X_1 = X_0^2 + \epsilon_1, & X_2 = \sin(X_0) + \epsilon_2, & X_3 = 0.5 X_0 + \epsilon_3, \\ X_4 \sim Unif(0, 10), & X_5 = 2\cos(\theta) + \epsilon_5, & X_6 = 2\sin(\theta) + \epsilon_6, & \end{array} \tag{4.1}$$

where

$$\epsilon_1 \sim \mathcal{N}(0,1),\ \epsilon_2 \sim \mathcal{N}(0, \frac{1}{4^2}),\ \epsilon_3 \sim \mathcal{N}(0, \frac{1}{4^2}),\ \epsilon_5 \sim \mathcal{N}(0, \frac{1}{10^2}), \epsilon_6 \sim \mathcal{N}(0, \frac{1}{10^2}),\ \theta \sim Unif(0, 2\pi).$$

By construction, there are three independent groups of variables in the model (4.1)

$$X_{S_1} = (X_0, X_1, X_2, X_3), \quad X_{S_2} = X_4, \quad X_{S_3} = (X_5, X_6), \tag{4.2}$$

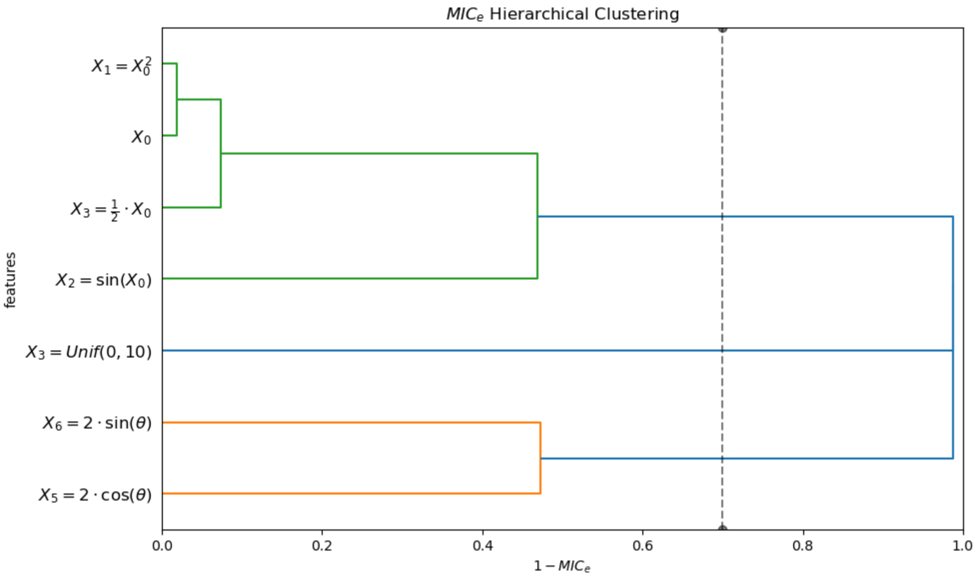
such that within each group the variables have strong dependencies. Figure 5 displays scatter plots of 10^4 samples of paired variables from the joint distribution (4.1) that visually confirms the grouping (4.2).

Figure 4a displays a dendrogram generated by the MIC$_e$-based dissimilarity measure, whose geometry is in accordance with our intuition on how predictors should be grouped with each other based on their dependencies and the accompanying noise level. Using the dendrogram as a guide, setting the dissimilarity threshold $\alpha = 0.7 \geq 1 - \text{MIC}_e$, we conclude that the variables are partitioned into groups $\mathcal{P}_{\alpha=0.7}^{\text{MIC}_e} = \{S_1, S_2, S_3\}$ with S_i given by (4.2), which coincides with the built-in grouping.

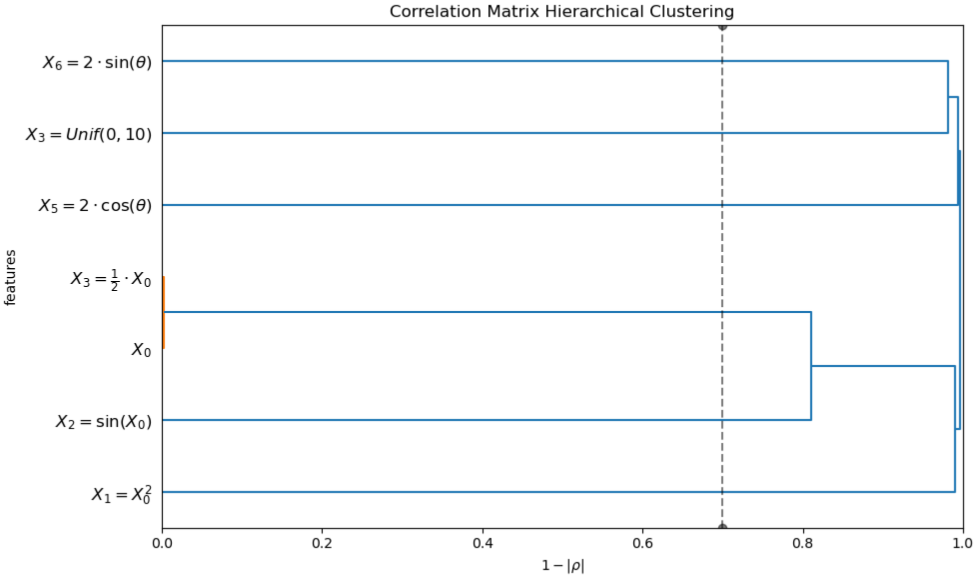
In contrast, according to the dendrogram on Figure 4b, the correlation-based clustering fails to capture non-linear dependencies as it ignores the sine functional dependence and captures weak dependencies between X_5 and X_6 that form a noisy circle, placing them in different clusters. Setting the dissimilarity threshold $\alpha = 0.7 \geq 1 - |\rho|$ with ρ the Pearson correlation, we obtain $\mathcal{P}_{\alpha=0.7}^{\rho} = \{\{0,3\}, \{1\}, \{2\}, \{4\}, \{5\}, \{6\}\}\}$, which is drastically different from the designed grouping (4.2).

5 Numerical examples

This section contains examples that provide numerical evidence for the stability bounds discussed in §3. The theoretical results from that section show that the bound for marginal explanations in $L^2(P_X)$ becomes larger as the dependencies between features become stronger (see Propositions 3.4 and 3.5), which in turn makes the explanations more unstable. Example 3.1, on the other hand, demonstrates that marginal explanations may not even be continuous in $L^2(P_X)$; for an extensive stability analysis of single feature explanations see the companion paper [45].

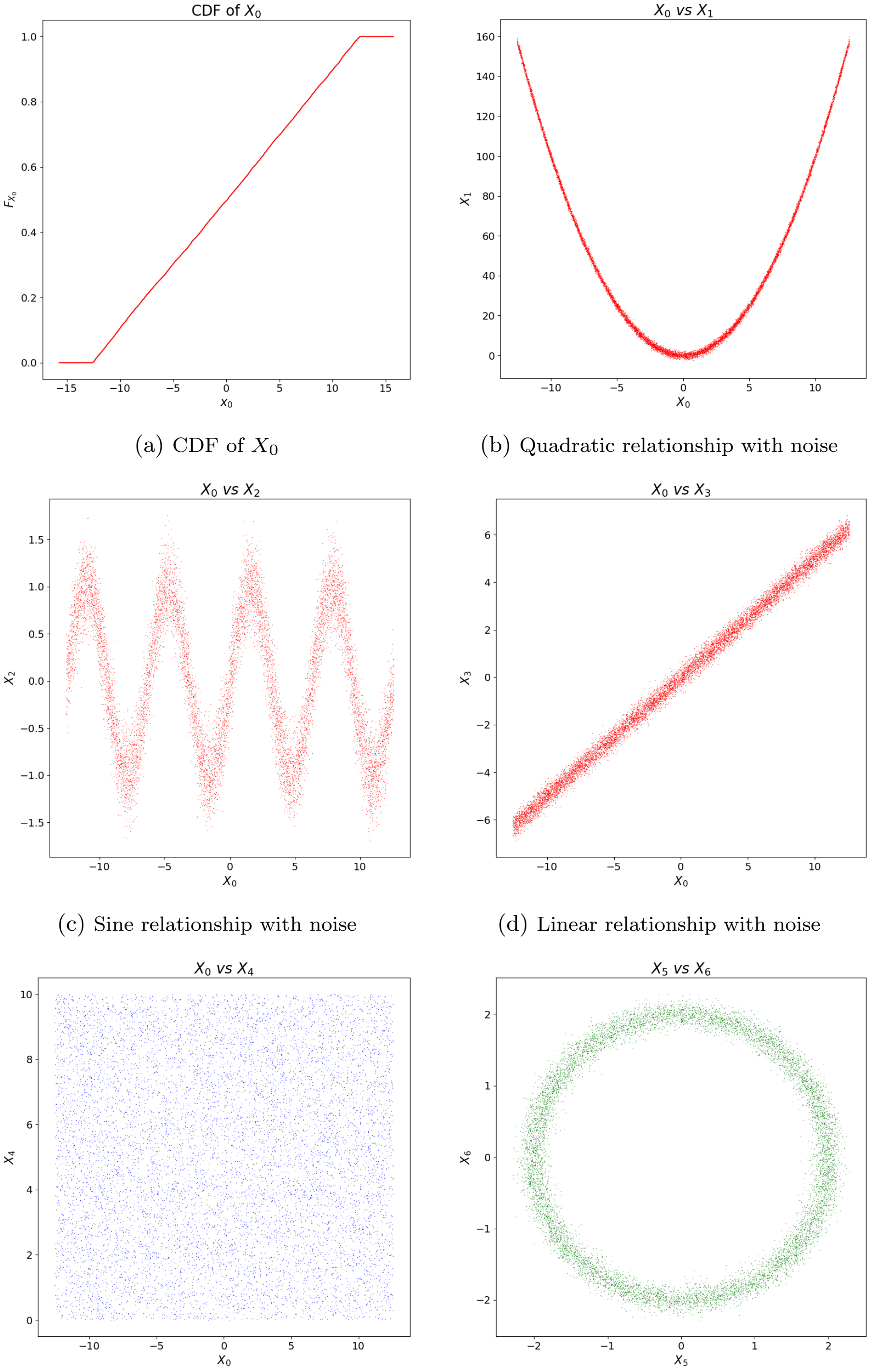


(a) MIC-based hierarchical clustering with GA linkage.



(b) Correlation-based hierarchical clustering with GA linkage.

Figure 4: Variable hierarchical clustering for the model (4.1). The dotted vertical line is based on a dissimilarity threshold; the predictors that have remained together on the left of it end up in same groups.



(a) CDF of X_0

(b) Quadratic relationship with noise

(c) Sine relationship with noise

(d) Linear relationship with noise

(e) Independent relationship

(f) circle

Figure 5: Scatter plots showing the dependencies in the distribution of (4.1).

Demonstrating explanation instability in $L^2(P_X)$ numerically is not a trivial task because the space $L^2(P_X)$ of models is much larger than any class of models obtained via training. Nevertheless, we numerically investigate the stability of marginal group explanations by first constructing a collection of "similar" ML models and then comparing the differences between the resulting single feature explanations with the differences between predictions of those models (which provide us with an estimate of the stability bounds). We then repeat the analysis for groups of features, where grouping is done by means of hierarchical clustering based on dependencies. In what follows, we first explore models built for a synthetic dataset and then a public real-world dataset.

5.1 Synthetic example on instability of marginal explanations

We consider the following data generating model[3]. Let $X = (X_1, X_2, X_3)$ be predictors such that the pair (X_1, X_2) is independent of X_3, with the distribution given by

$$\begin{aligned} &Z \sim Unif(-1,1), \quad X_1 = Z + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0,\delta), \\ &X_2 = \sqrt{2}\sin(Z(\pi/4)) + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0,\delta), \quad X_3 \sim Unif\big([-1,-0.5] \cup [0.5,1]\big), \end{aligned} \tag{5.1}$$

where $\delta > 0$ is chosen later. The model for the output variable is assumed to be

$$Y = f_*(X_1, X_2, X_3) = 3X_2X_3. \tag{5.2}$$

Note that in the true regressor f_* the variable X_1 is a dummy variable (not explicitly used). For this reason, the marginal explanation approach will assign zero attribution in f_* to this variable.

The dependencies in predictors allow for the existence of many models with distinct representations from $L^2(P_X)$ that approximate the response variable well. In what follows, we demonstrate that the generated explanations differ between such models. We then demonstrate how grouping features by dependencies increases stability or, more precisely, how it reduces the stability bounds in $L^2(P_X)$.

Models on perturbed datasets. In this experiment, we construct five distinct datasets by varying the level of noise in the predictors from the previous subsection, and train five corresponding ML models. We then construct a test dataset as a mixture of the five training sets and use its observations for both explanations and averaging. This experiment demonstrates that the models with similar predictive power on the test dataset, which in turn is close in distribution to the training sets, have widely different explanations and how grouping features based on dependencies mitigates the explanation instabilities.

First, for each $\delta \in \{\delta_i\}_{i=1}^5 = \{0.0, 0.001, 0.0025, 0.005, 0.01\}$, which represents the noise level in predictors of the data-generating model (5.1), we construct a corresponding dataset $D(\delta) = \{(x_\delta^{(k)}, y_\delta^{(k)})\}_{k=1}^K$, containing $K = 25000$ observations sampled from the distribution (X_δ, Y_δ) where X_δ is given by (5.1) with noise $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0,\delta)$, and Y_δ is constructed using the response model (5.2). Then for each $i \in \{1,\dots,5\}$ an XGBoost regressor $f_i(x)$ is trained on the dataset $D(\delta_i)$, utilizing the following hyperparameters[4]: n_estimators=300, max_depth=5, subsample=1.0, learning_rate=0.1, alpha=10, lambda=10.

To compare the explanations of these models, a common test dataset $D = \{(x^{(k)}, y^{(k)})\}_{k=1}^K$ is constructed by drawing $K = 25000$ samples from the distribution (X, Y) such that $X = \sum_{i=1}^5 1_{\{C=i\}} \cdot X_{\delta_i}$ is a mixture, where C is a random variable satisfying $\mathbb{P}(C = i) = 0.2$, and Y is obtained using the response model (5.2).

Performance metrics for the XGBoost models on the mixture dataset were evaluated. Specifically, the relative L^2-errors for the five models are approximately 0.051, 0.045, 0.041, 0.052 and 0.046, respectively, with the norms $\|f_i\|_{L^2(P_X)}$ recorded in Table 1, illustrating that all trained models have similar predictive power on the test set.

We next evaluate the $L^2(P_X)$-distance between the true model f_* and each trained model f_k, $k \in \{1,\dots,5\}$. The estimated values are recorded in Table 1 and showcase that the predictions of the trained

[3] Part of this example appears in the supplementary material of [45], where only single feature explanations are presented. Here, we do the analysis for both single and group level features to understand the effects of grouping.

[4] The code is available at link.

	$\|\cdot\|$	β_1	β_2	β_3	$\|\beta\|$	$\beta_1^{\mathcal{P}}$	$\beta_2^{\mathcal{P}}$	$\|\beta^{\mathcal{P}}\|$
f_1	1.370	0.674	0.000	0.675	0.954	0.674	0.675	0.954
f_2	1.375	0.334	0.385	0.685	0.854	0.681	0.683	0.964
f_3	1.374	0.285	0.461	0.682	0.871	0.680	0.681	0.962
f_4	1.375	0.214	0.228	0.040	0.315	0.679	0.681	0.962
f_5	1.374	0.068	0.627	0.677	0.925	0.678	0.680	0.960
f_*	1.380	0.000	0.682	0.682	0.965	0.682	0.683	0.965
$f_1 - f_*$	0.069	0.674	0.682	0.036	0.960	0.038	0.036	0.052
$f_2 - f_*$	0.062	0.334	0.336	0.035	0.475	0.032	0.028	0.042
$f_3 - f_*$	0.056	0.284	0.288	0.033	0.406	0.029	0.026	0.039
$f_4 - f_*$	0.071	0.214	0.228	0.041	0.315	0.044	0.029	0.053
$f_5 - f_*$	0.064	0.067	0.075	0.029	0.104	0.027	0.026	0.037

Table 1: Global marginal Shapley attributions.

models on the mixture dataset are close in an L^2-sense to those of f_*. This implies that $\{f_k\}_{k=1}^5$ are in an (L^2, ϵ)-Rashomon set of models about f_* (defined in §2.1) with $\epsilon = 0.071$, which constitutes about 5% relative L^2-distance.

We next pick $m = 1000$ samples at random from the mixture dataset, to construct the dataset $D_X^{(e)}$ of predictor observations used for explanations. We also subsample the predictors from the mixture set and obtain a background dataset $\bar{D}_X$ with 1000 samples used for constructing the empirical marginal game defined by

$$\hat{v}^{\scriptscriptstyle ME}(S; x, f, \bar{D}_X) := \frac{1}{|\bar{D}_X|} \sum_{\tilde{x} \in \bar{D}_X} f(x_S, \tilde{x}_{-S}) \approx \mathbb{E}[f(x_S, X_{-S})], \quad \text{for } x \in D_X^{(e)}. \tag{5.3}$$

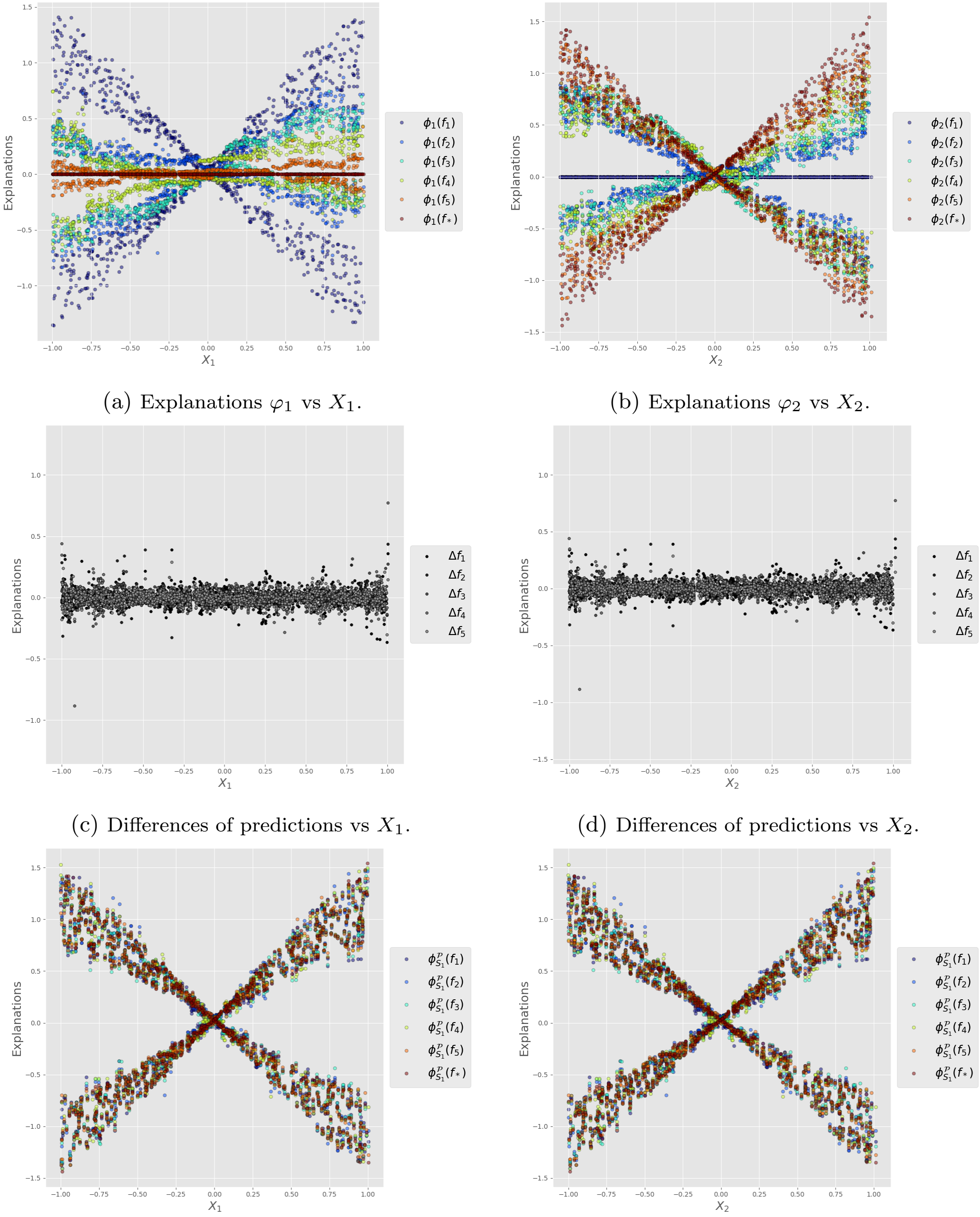
We then evaluate the empirical marginal explanations $\varphi_i[N, \hat{v}^{\scriptscriptstyle ME}](x)$ for each observation $x \in D_X^{(e)}$ and each predictor across the six models, the true model and the five XGBoost models. The computations are done by means of the interventional TreeSHAP method [42]. Figures 6a-6b depict the scatterplots of explanations for each model across the dataset $D_X^{(e)}$, where we see that explanations differ substantially, indicating the different functional representations.

To quantify the global attribution of each predictor, we estimate the L^2-norms of the marginal Shapley values for each model, $\beta_i(f_k, \hat{v}^{\scriptscriptstyle ME}) := \|\varphi_i(X; f_k, \hat{v}^{\scriptscriptstyle ME})\|_{L^2(\mathbb{P})}$, $i \in N$, which are depicted in Figure 7a and recorded in Table 1. These values also demonstrate that the features X_1, X_2 are utilized differently across the models.

Recall that by Proposition 3.1(iii) (due to the efficiency property of φ) the conditional Shapley operator is a linear, bounded operator with norm bounded by one and, hence, the conditional Shapley value satisfies $|\beta(f_1 - f_2, v^{\scriptscriptstyle CE})|/\|f_1 - f_2\|_{L^2(P_X)} \leq 1$, where $\beta := (\beta_1, \beta_2, \beta_3)$. This bound ensures that the total distance $|\beta(f_1 - f_2, v^{\scriptscriptstyle CE})|$ between these explanations is always smaller than the $L^2(P_X)$-distance between the models, and the same is true for any component and sub-vector of $\beta(f_1 - f_2, v^{\scriptscriptstyle CE})$. Meanwhile, in the presence of dependencies, the bound for the marginal explanations may in general be significantly larger than one, depending on the relationship between P_X and $\widetilde{P}_X$.

To assess the degree of the instability in marginal explanations, we estimate the distance between the marginal Shapley values of the reference model f_* and f_k for every $k \in \{1, \ldots, 5\}$ and each predictor $X_i, i \in \{1, 2, 3\}$, which are equal to the norm of the Shapley values for the model difference $\beta_i(f_i - f_*, \hat{v}^{\scriptscriptstyle ME}) = \|\varphi_i(X; f_i - f_*, \hat{v}^{\scriptscriptstyle ME})\|_{L^2(\mathbb{P})}$, and then compare those with the model distances; see Figure 7b, where $f_k - f_*$ is denoted as Δf_k, showcasing the comparison for the single feature explanations.

We contrast the unit operator bound in (3.3) for conditional explanations in relation to the change in empirical marginal explanations with respect to the $L^2(P_X)$-distance between models. Specifically, the ratio of the marginal explanation distance to the distance between models varies from approximately 1 to 10; see



(a) Explanations φ_1 vs X_1.

(b) Explanations φ_2 vs X_2.

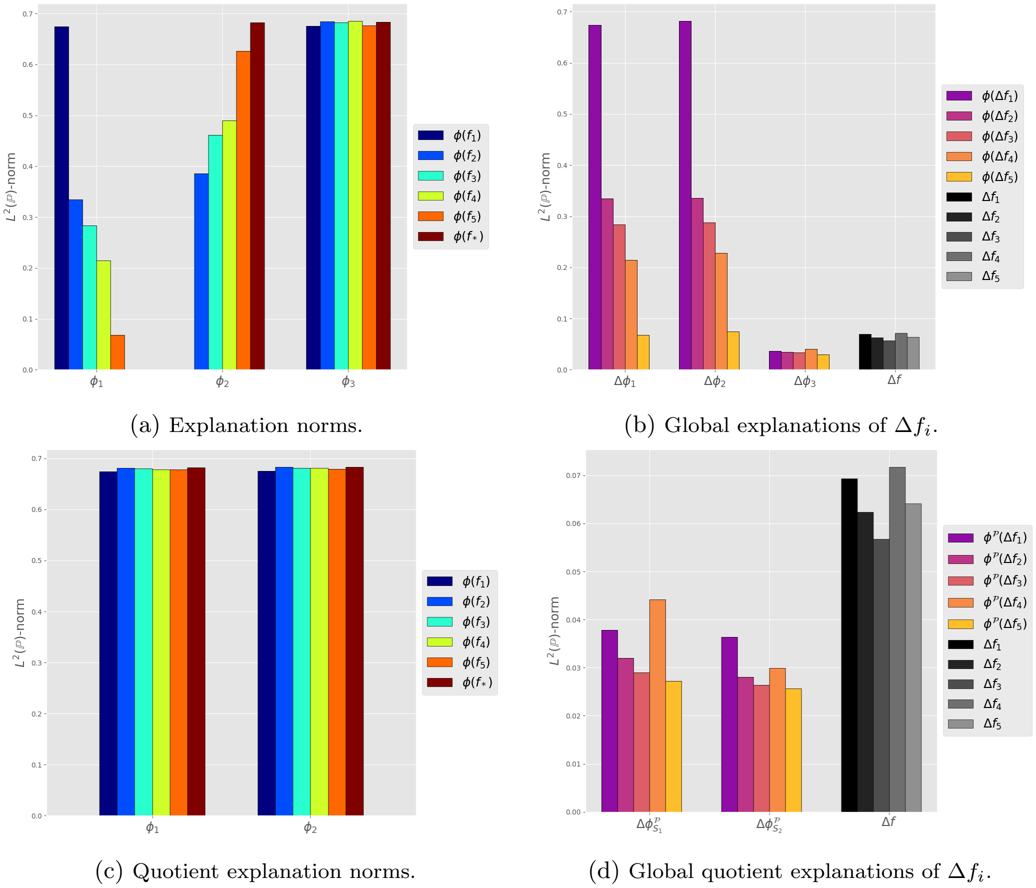
(c) Differences of predictions vs X_1.

(d) Differences of predictions vs X_2.

(e) Explanations $\varphi^{\mathcal{P}}_{S_1}$ vs X_1.

(f) Explanations $\varphi^{\mathcal{P}}_{S_2}$ vs X_2.

Figure 6: Individual and quotient marginal explanations.



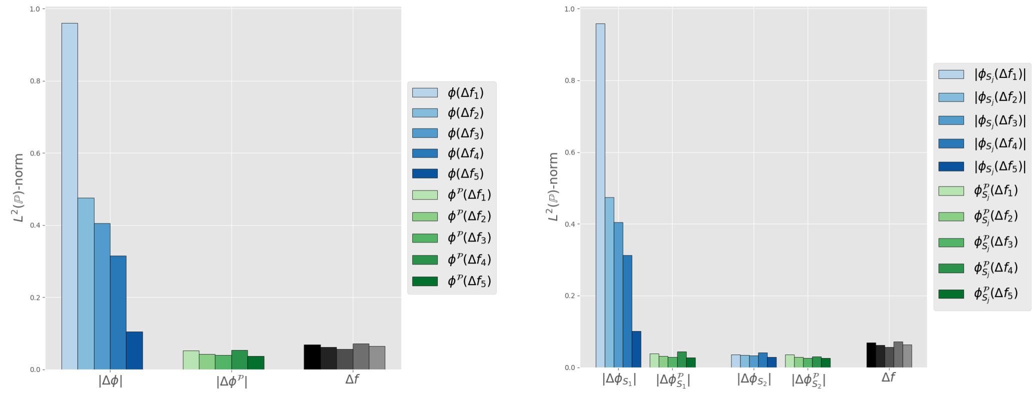
(a) Explanation norms.

(b) Global explanations of Δf_i.

(c) Quotient explanation norms.

(d) Global quotient explanations of Δf_i.

Figure 7: Individual and quotient explanation norms.



(a) Total gain in stability.

(b) Gain in stability.

Figure 8: Explanation norms and effect of grouping.

Figure 7b. Note that the differences between explanations are significant and for some models constitute about 50% of the true model's norm. Observe also that the total distances between the vectors of global marginal explanations satisfy $\{|\beta(f_i - f_*, \hat{v}^{ME})|\}_{i=1}^5 = \{0.960, 0.475, 0.406, 0.315, 0.104\}$ and are approximately two-to-fourteen times larger than the $L^2(P_X)$-distance between models.

To understand the effect of grouping, we next construct quotient marginal explanations of the trained models for each sample $x \in D_X^{(e)}$. To accomplish this, we employ the empirical quotient marginal game and generate, via direct computation, quotient Shapley values $\varphi_j[M, \hat{v}^{ME,\mathcal{P}}](x)$, $j \in M = \{1,2\}$, for each $x \in D_X^{(e)}$ corresponding to the partition $\mathcal{P} = \{\{1,2\},\{3\}\} = \{S_1, S_2\}$. Figures 6e-6f depict the scatterplots of quotient explanations for each model across the dataset D_X, where we see that the explanations between the models are similar.

We then use these explanations to quantify the global attribution of predictor groups by estimating the norms $\beta_j^{\mathcal{P}}(f_k, \hat{v}^{ME}) := \|\varphi_{S_j}^{\mathcal{P}}(X; f_k, \hat{v}^{ME})\|_{L^2(\mathbb{P})}$, $j \in M$, which are depicted in Figure 7c and recorded in Table 1. These values indicate that grouping by dependencies yields (on average) similar group explanations regardless of the functional representation.

To quantify the difference between quotient explanations, we estimate the $L^2(P_X)$-distances between marginal quotient explanations for the partition $\mathcal{P}$, denoted by $\beta_j^{\mathcal{P}}(f_i - f_*, \hat{v}^{ME})$, $j \in \{1,2\}$, respectively. Figure 7d compares the latter with the distances between the models as given in Table 1. We see that these distances (due to grouping) are approximately twice smaller than the distances between the models compared to individual explanations, showcasing the consistency with the bound for conditional explanations.

To estimate the gain in stability due to grouping, we apply the following approach that will be useful when dealing with large datasets where dependencies are not always that obvious. Specifically, we compare the norm of explanation vectors $|\beta(f_i - f_*, \hat{v}^{ME})|$ and $|\beta^{\mathcal{P}}(f_i - f_*, \hat{v}^{ME})|$ to quantify the total gain in stability (across all features simultaneously), which is depicted in Figure 8a. We also compare the norms of the quotient explanations' differences for each $j \in M$ with the length of corresponding subvectors $|\beta_{S_j}(f_i - f_*, \hat{v}^{ME})|$. Figure 8b illustrates that the differences in aggregated individual explanations drop significantly after grouping, and well below the $L^2(P_X)$-norm of the model difference, which showcases the gain in stability across each group.

5.2 Experiments with public datasets

5.2.1 Default of Credit Card Clients

In this section, we apply the group explanation techniques to public datasets. We start our investigation with the Default of Credit Card Clients dataset [66] from the UCI Machine Learning Repository. This dataset contains 30000 instances, 23 features and a dependent binary variable Y that indicates if an individual defaulted on a payment, where the default is denoted by $Y = 1$. The protected attributes 'sex', 'marriage', and 'age' were removed in order to be consistent with regulatory practices. The remaining twenty predictors were used for model training, where we use the training dataset D_{train} with 27000 samples to build a classification score $p_*(x) := \widehat{\mathbb{P}}(Y = 1|X = x)$ using the CatBoost algorithm, whose corresponding population minimizer is defined by $f_*(x) = \text{logit}(p_*(x))$. For training we use the following parameters: iterations=200, min_data_in_leaf=5, depth=5, subsample=0.8, and learning_rate=0.1.

Performance metrics for the model on the trained dataset, and test dataset with 3000 samples, were evaluated. Specifically, the mean logloss on the train and test set is approximately 0.40 and 0.41 respectively, and the AUC is 0.82 and 0.80 respectively.

To assess the dependencies, we build a dendrogram based on the MIC-metric and investigate the level of dependence that exists among the twenty predictors. As we will see, the Rashomon effect is still present in the models trained on the Default of Credit Clients dataset and grouping leads to improved stability, although not as drastic given the lack of strong dependencies.

Grouping effect on stability. First, we explore the Rashomon effect by measuring and comparing the stability of explanations before and after grouping on the Default of Credit Card Clients dataset [66]. To understand how the dependencies between groups affect stability, we design the following experiment. Given the reference model f_* and the population minimizer described in the beginning of Section 5.2, we train a

	$\|f_*\|$	$\max_k \|\Delta f_k\|$	$\max_k \lvert\beta(\Delta f_k, \mathcal{P})\rvert$	$\max_k \lvert\beta^{\mathcal{P}}(\Delta f_k, \mathcal{P})\rvert$
$\mathcal{P}_{0.49}$	1.839	0.267	0.388	0.364
$\mathcal{P}_{0.62}$	1.839	0.267	0.390	0.388
$\mathcal{P}_{0.65}$	1.839	0.267	0.389	0.340
$\mathcal{P}_{0.77}$	1.839	0.267	0.386	0.269

Table 2: Global marginal Owen attributions for Default of Credit Card Clients dataset.

series of new models whose predictions are close to those of f_* by varying the hyperparameters. Specifically, we pick the following parameters at random from the given intervals: iterations $\in [50, 300]$, subsample $\in [0.5, 1.0]$, depth $\in [2, 10]$, learning_rate $\in [0.025, 0.25]$, rsm $\in [0.5, 0.1]$, with the rest of the parameters being the same as for f_*. We then train a new model f and accept or reject based on the following principle. Given a threshold $\tau \in [0, 1]$, we accept the model if it is in the Rashomon ball of relative radius τ, meaning that $\hat{\mathbb{E}}[|f(X) - f_*(X)|^2] \leq \tau \|f_*\|_{L^2(P_X)}$, otherwise it is rejected; here $\|f_*\|_{L^2(P_X)} \approx 1.84$ which is estimated on the test set, and we choose $\tau = 0.1$. We continue this procedure until we train 20 models $\{f_k\}_{k=1}^{20}$.

We next compute the Owen values (see (3.18)) $\{Ow(x; \hat{v}^{ME}, \mathcal{P}, f_*)\}_{i=1}^{20}$ of the empirical marginal game for different partitions $\mathcal{P}$ obtained by thresholding the variable clustering tree, depicted on Figure 11, based on MIC_e. Note that the tree can be viewed as a coalescent tree parameterized by $\alpha = 1 - \mathrm{MIC}_e$. This yields a sequence of nested partitions $\{\mathcal{P}^{(k)}\}_{k=0}^{19}$, with $\mathcal{P}^{(k)}$ corresponding to the k-th coalescent, and having $(20 - k)$ groups of predictors.

In particular, we consider partitions of features with varying degrees of dependence by cross-sectioning the partition tree of dependencies for a given dataset at different heights. In the analysis that follows the partitions considered are $\mathcal{P}_{0.49}$, $\mathcal{P}_{0.62}$, $\mathcal{P}_{0.65}$ and $\mathcal{P}_{0.77}$, containing 12, 10, 9, and 5 groups (see Figure 9), respectively, with the subscript indicating the cutoff threshold.

To accomplish this, we use the empirical game defined in (5.3) with a background dataset $\bar{D}_X := D_{train}$. To compute the explanations of f_*, given the dimensions of the background dataset, we use the fast, exact algorithm introduced in Filom et al. [20], which is designed specifically for the computation of empirical marginal coalitional values of CatBoost ensembles.

Similarly, given these four partitions we evaluate the empirical marginal Owen explanations for each predictor and for each group (via summation) per model $\{f_k\}_{k=1}^{20}$. Subsequently, we compute the global explanation of the model difference $\|f_* - f_k\|_{L^2(\mathbb{P})}$ from individual explanations. Specifically, we first evaluate the explanations $Ow_i(X, f_* - f_k, \hat{v}^{ME}, \mathcal{P})$ of the model difference $\Delta f_k = f_* - f_k$ and then compute the corresponding norms $\beta_i(f_* - f_k, \mathcal{P}) = \|Ow_i(x, f_* - f_k, \mathcal{P})\|_{L^2(\mathbb{P})}$, $i \in N$, for each $k \in \{1, \ldots, 20\}$ and $\mathcal{P} \in \{P_{0.49}, P_{0.62}, P_{0.65}, P_{0.77}\}$. We do the same for the group explanations, $\beta_j^{\mathcal{P}}(f_k - f_*, \mathcal{P}) = \|Ow_{S_j}(X, f_k - f_*, \hat{v}^{ME}, \mathcal{P})\|_{L^2(\mathbb{P})}$, $j \in M$.

To contrast the stabilization effect between individual and group explanations, we evaluate the length of the vectors $\beta(f_* - f_k, \mathcal{P}) = \{\beta_i(f_* - f_k, \mathcal{P})\}_{i \in N}$ and $\beta^{\mathcal{P}}(f_* - f_k, \mathcal{P}) = \{\beta_j^{\mathcal{P}}(f_* - f_k, \mathcal{P})\}_{j \in M}$ and compute the maximum of these quantities across all models $k \in \{1, \ldots, 20\}$. These are plotted in Figure 10a together with the norm of the maximum model difference. As we see in the plot, the total group explanation differences are smaller than the respective total individual ones, showcasing the gain in stability when considering explanations of groups. Furthermore, as we reach partition $\mathcal{P}_{0.77}$, observe that the total group explanation difference becomes approximately equal to the norm of the difference of the models, illustrating the alleviation of the Rashomon effect due to weaker dependencies between the groups. Note that in Figure 10b we have removed the energy contributed by singletons from both individual and group explanation vectors, since these do not have any effect on the norm evaluation between the two.

Given that the dependencies are not very strong in the Default of Credit Clients dataset, as seen in the partition tree in Figure 9, note that the Rashomon effect in general is not as prominent as in the synthetic example from §5.1, but it is still present. Nevertheless, Figure 10 still portrays this effect and its alleviation when predictor groups are considered based on dependencies.

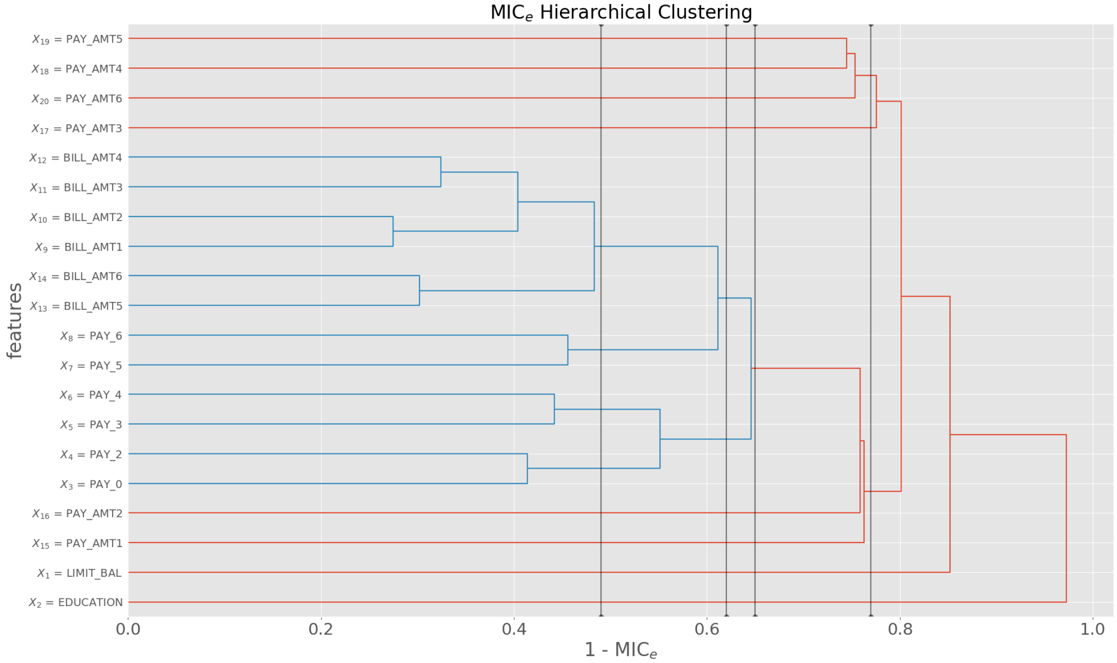
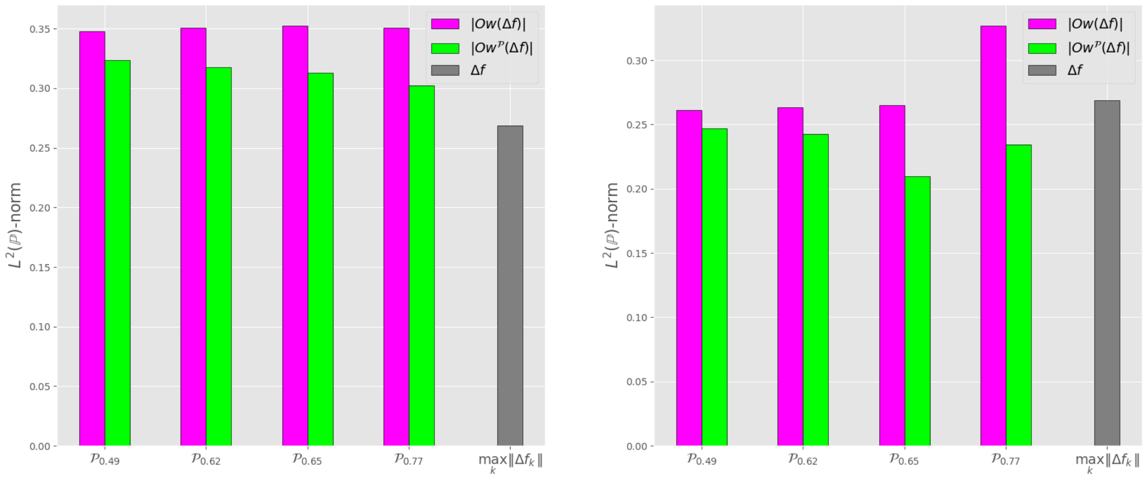


Figure 9: MIC-based hierarchical clustering for the Default of Credit Card Clients dataset with 4 cutoffs.



(a) Norms of explanation differences (with singletons).

(b) Norms of explanation differences (without singletons).

Figure 10: Gain in stability, $|\beta|$ versus $|\beta^{\mathcal{P}}|$. Default of Credit Card Clients dataset.

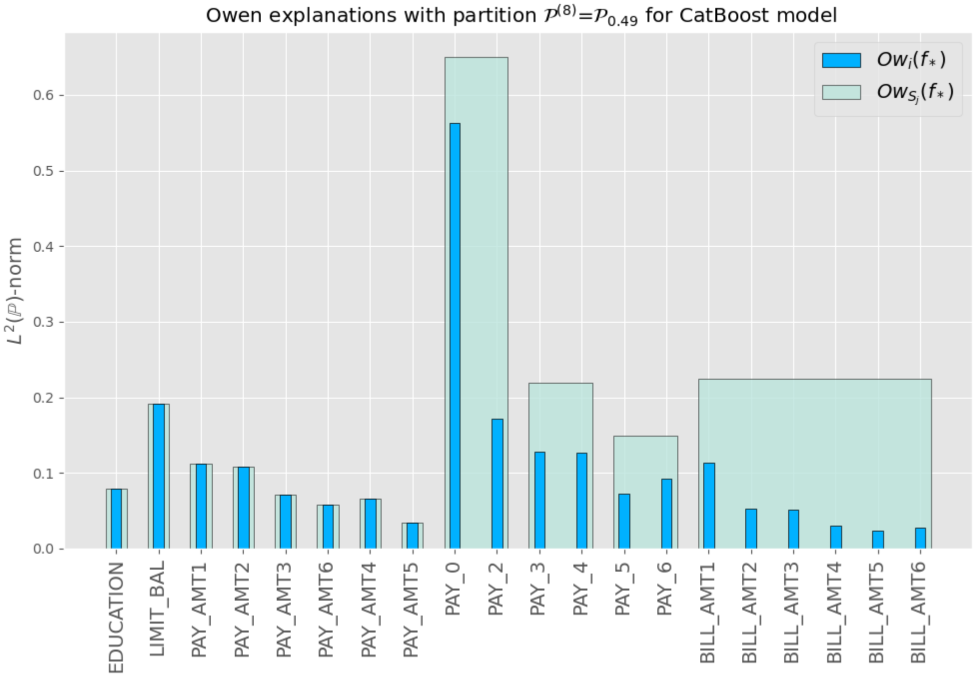


Figure 11: Global Owen explanation for $\mathcal{P}^{(8)} = \mathcal{P}_{0.49}$.

Grouping effect on ranking. In the companion paper [45, Section 3.3] we have shown that under dependencies in predictors the marginal explanations have the explanations (or more strictly their energy) split among dependent predictors; see also [1]. In what follows, we illustrate how grouping helps to mitigate the splitting among dependencies and how it impacts the ranking; see also the analysis in [34].

Consider the partition $\mathcal{P}^{(8)} = \mathcal{P}_{0.49}$ that can be obtained by thresholding the dendrogram, depicted on Figure 9, at $\alpha = 0.49$. This partition contains the following groups: {PAY_0,PAY_2}, {PAY_3,PAY_4}, {PAY_5,PAY_6}, {BILL_AMT1,...,BILL_AMT6}, and the rest are singletons.

Next, we compute the (global) empirical marginal Owen explanations of the population minimizer $f_*(x)$, the values $\beta_i(f_*, \hat{v}^{ME}) = \|Ow_i(X; \hat{v}^{ME}, \mathcal{P}^{(8)}, f_*)\|_{L^2(\mathbb{P})}$, $i \in N$, which are depicted in Figure 11. As before, we use the empirical game defined in (5.3) with a background dataset $\bar{D}_X := D_{train}$ and compute the explanations of the population minimizer using the algorithm [20].

We then compute the trivial group explanations (see Definition 3.7) over each group to obtain the global contributions of the groups themselves, that is, the values $\beta_j^{\mathcal{P}^{(8)}}(f_*, \hat{v}^{ME}) = \|Ow_{S_j}(X; \hat{v}^{ME}, \mathcal{P}^{(8)}, f_*)\|_{L^2(\mathbb{P})}$, $S_j \in \mathcal{P}^{(8)}$, which are depicted in Figure 11. Recall that the group sums of Owen values are equal to the quotient Shapley values in view of the quotient game property (QP). Since groups are not fully independent, the quotient marginal Shapley values are only crude approximants of the conditional ones.

To observe the splits, it is sufficient to compare the contributions of highly dependent predictors that form the coalition with that of the coalition itself as well as with contributions of independent (or almost independent) predictors that form singletons and whose marginal explanations, according to Proposition 3.5, are equal to (or approximate well) the corresponding conditional ones.

The splits are prominent in Figure 11 which presents the norms of contributions of the individual predictors together with the corresponding groups. Observe the energy splitting occurring in the predictor group with BILL_AMT's and contrast the individual and group explanations with, for example, the explanation of LIMIT_BAL. When one attempts to rank order predictors based on their contributions, LIMIT_BAL will be placed higher in the ranking compared to each BILL_AMT. However, when ranking groups, {LIMIT_BAL}, as a singleton, will be placed lower than the group containing the BILL_AMT predictors. This clearly

	$\|f_*\|$	$\max_k \|\Delta f_k\|$	$\max_k \lvert\beta(\Delta f_k, \mathcal{P})\rvert$	$\max_k \lvert\beta^{\mathcal{P}}(\Delta f_k, \mathcal{P})\rvert$
$\mathcal{P}_{0.3}$	47.482	2.808	11.286	7.207
$\mathcal{P}_{0.4}$	47.482	2.808	11.269	6.479
$\mathcal{P}_{0.5}$	47.482	2.808	11.119	5.018
$\mathcal{P}_{0.55}$	47.482	2.808	11.146	4.971
$\mathcal{P}_{0.60}$	47.482	2.808	11.150	4.054
$\mathcal{P}_{0.66}$	47.482	2.808	11.167	2.454

Table 3: Global marginal Owen attributions for Superconductivity dataset.

indicates the issue caused by energy splits to rank ordering based on contributions of individual predictors.

5.2.2 Superconductivity

We next consider the Superconductivity dataset [27], a regression dataset where the superconductivity critical temperature is predicted based on 81 features extracted from the superconductor's chemical formula. The original dataset has 21263 instances. As before, we first construct a hierarchical clustering tree of feature dependencies using the MIC-based metric (see Figure 12) in order to form partitions. The dataset is then randomly split into training and test sets in 90:10 proportions, and we train a (reference) regressor model $f_*(x) = \hat{\mathbb{E}}[Y|X = x]$ using the CatBoost algorithm. For training we use the following parameters: iterations=300, min_data_in_leaf=5, depth=8, subsample=0.8, and learning_rate=0.1.

Performance metrics for the model on the trained and test datasets, the latter with 2126 samples, were evaluated. Specifically, the mean square error estimate on the training and test sets is approximately 7.70 and 9.45 respectively, which constitutes about 16% and 20% of relative error given that the L^2-norm estimate of the reference model is $\|f_*\|_{L^2(P_X)} \approx 47.48$ on the test dataset.

Following the above methodology, we train a series of new models whose predictions are close to the predictions of f_*. Specifically, we pick the following parameters at random from the given intervals: iterations $\in [100, 500]$, subsample $\in [0.5, 1.0]$, depth $\in [4, 10]$, learning_rate $\in [0.025, 0.25]$, rsm $\in [0.5, 0.1]$, with the rest of the parameters being the same as for f_*. We then train a new model f and accept it if it is in the Rashomon ball centered at f_* of relative size $\tau = 0.06$, meaning if $\hat{\mathbb{E}}[|f(X) - f_*(X)|^2] \leq \tau\|f_*\|_{L^2(P_X)}$, or reject otherwise. We continue this procedure until we construct 25 models $\{f_k\}_{k=1}^{25}$.

Similar to the previous dataset, the partitions considered in this analysis are $\mathcal{P}_{0.3}$, $\mathcal{P}_{0.4}$, $\mathcal{P}_{0.5}$, $\mathcal{P}_{0.60}$ and $\mathcal{P}_{0.65}$; see Figure 12. Given these six partitions we evaluate, as before, the empirical marginal Owen explanations for each predictor and for each group (via summation) per model and then evaluate the length of their global explanations. These are plotted in Figure 13a together with the norm of the maximum model difference. Once again, the total group explanation differences are smaller than the respective total individual ones, showcasing the gain in stability when considering explanations of groups. Furthermore, as we reach partition $\mathcal{P}_{0.65}$, observe that the total group explanation difference becomes approximately equal to the norm of the difference of the models, illustrating again the alleviation of the Rashomon effect.

We would like to contrast this dataset with the Default of Credit Card Clients dataset. Note that due to the stronger dependencies among the features of the Superconductivity dataset, the Rashomon effect is much more apparent in this case compared to the previous dataset, which also means that the alleviation of the Rashomon effect due to evaluating group explanations is also more striking.

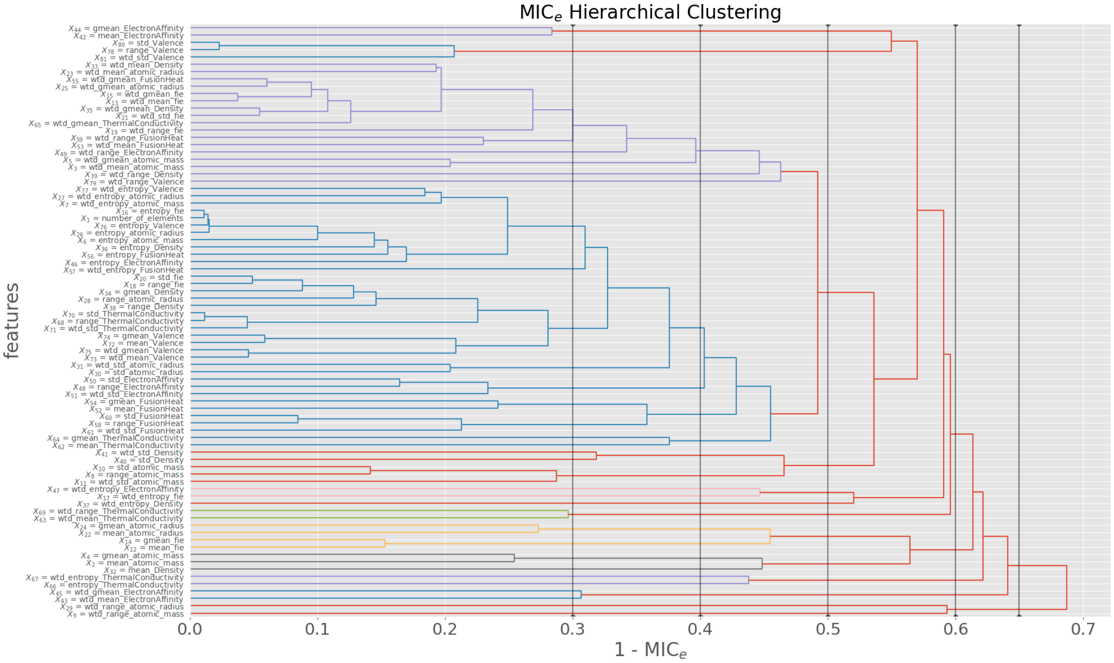
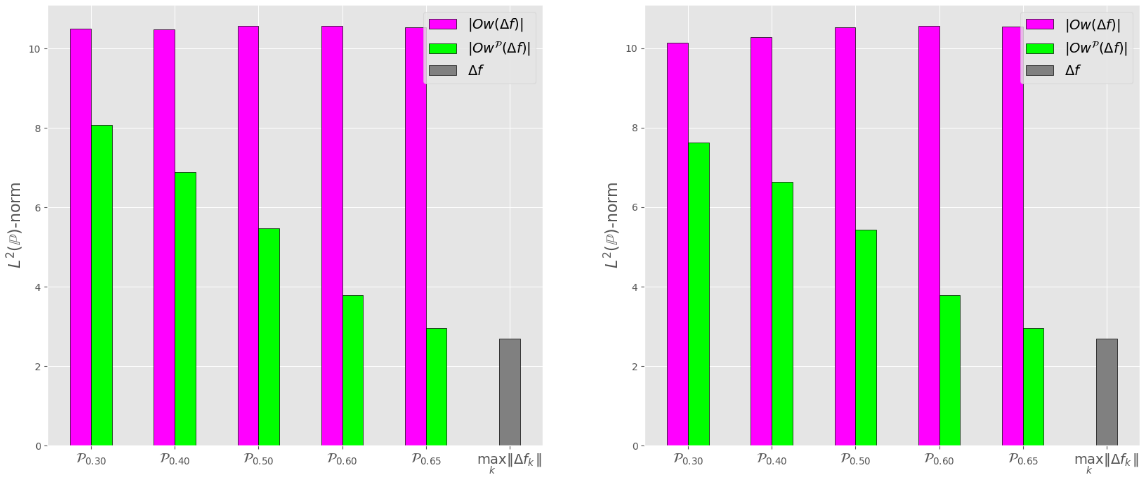


Figure 12: MIC-based hierarchical clustering for the Superconductivity dataset with 6 cutoffs.



(a) Norms of explanation differences (with singletons).

(b) Norms of explanation differences (without singletons).

Figure 13: Gain in stability, $|\beta|$ versus $|\beta^{\mathcal{P}}|$. Superconductivity dataset.

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, $N = \{1, 2, \dots, n\}$) 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 $w(S, n) \geq 0, S \subset N$.

B Two-step formulation for coalitional value

Here we establish that a coalitional value g^w defined in (3.17) can be expressed in terms of the linear games defined in (3.19).

Lemma B.1 (two-step representation). *Let g^w and the weights $w = (w^{(1)}, w^{(2)})$ be as in* (3.17)*. Let α_*, $h_*^{(1)}$, $h_*^{(2)}$ be induced by g^w as in Definition 3.6. Then:*

(i) For $i \in N$ we have

$$g_i^w[N, v, \bar{N}] = \alpha_* h_{*,i}^{(1)}[N, v], \quad g_i^w[N, v, \{N\}] = \alpha_* h_{*,}^{(2)}[N, v]. \tag{B.1}$$

(ii) For $j \in M$ and $i \in S_j \in \mathcal{P} = \{S_1, \dots, S_m\}$ we have

$$g_i[N, v, \mathcal{P}] = \alpha_* h_{*,i}^{(2)}[S_j, v^{(j)}], \quad v^{(j)}(T) := h_{*,j}^{(1)}[M, v^{\mathcal{P}_{|T}}], \quad T \subseteq S_j \tag{B.2}$$

where for each $T \subseteq S_j$ we have

$$v^{\mathcal{P}_{|T}}(A) := \mathbb{1}_{\{j \notin A\}} v^{\mathcal{P}}(A) + \mathbb{1}_{\{j \in A\}} v(\cup_{k \in A \setminus \{j\}} S_k \cup T), \ A \subseteq M. \tag{B.3}$$

Proof. Follows from the direct calculation. □

Let us next provide the two-step representation for the Owen and Banzhaf-Owen values. Specifically, for each $j \in M$, $i \in S_j$, and $T \subseteq S_j$ we have

$$\begin{aligned} Ow_i[N, v, \mathcal{P}] &= \varphi_i[S_j, v_{Ow}^{(j)}], & v_{Ow}^{(j)}(T) &:= \varphi_j[M, v^{\mathcal{P}_{|T}}] \\ BzOw_i[N, v, \mathcal{P}] &= Bz_i[S_j, v_{BzOw}^{(j)}], & v_{BzOw}^{(j)}(T) &:= Bz_j[M, v^{\mathcal{P}_{|T}}]. \end{aligned} \tag{B.4}$$

Notice that one has $\alpha_* = 1$ and that $h_*^{(1)} = h_*^{(2)} = \varphi$ (see (2.3)) in the first two-step formulation while $h_*^{(1)} = h_*^{(2)} = Bz$, where $Bz_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{2^{n-1}} \big[v(S \cup \{i\}) - v(S) \big]$, $i \in N$, is the well-known Banzhaf value [8], in the second.

Remark B.1. We note that the representation (B.2) of g^w is unique up to a scaling constant, and it is symmetric with respect to the partition elements $S_1, \cdots, S_m$.

C Proofs of the results from §3

C.1 Proof of Proposition 3.1

Proof. Combining (2.4) and Definition 2.2, we have $v^{CE,\mathcal{P}}(A; X, f) = \mathbb{E}[f(X)|X_{Q_A}]$, where $Q_A = \cup_{r \in A} S_r$. Here, the conditional expectation is a well-defined element of $L^2(\mathbb{P})$ since $f \in L^2(P_X)$, or equivalently, $f(X)$ belongs to space $L^2(\mathbb{P})$. To show the bound, we write

$$\|v^{CE,\mathcal{P}}(A; X, f)\|_{L^2(\mathbb{P})}^2 = \mathbb{E}[\mathbb{E}[f(X)|X_{Q_A}]^2] \leq \mathbb{E}[\mathbb{E}[f^2(X)|X_{Q_A}]] = \mathbb{E}[f^2(X)] = \|f(X)\|_{L^2(\mathbb{P})}^2 = \|f\|_{L^2(P_X)}^2,$$

where we used Jensen's inequality and the law of total expectation. This proves part (i).

As for part (ii), as h is of the form (3.1), the triangle inequality indicates that

$$\|\bar{\mathcal{E}}_j^{CE}[f; h, X, \mathcal{P}]\|_{L^2(\mathbb{P})} \leq \sum_{A \subseteq M \setminus \{j\}} w(A, m) \cdot \big\| \mathbb{E}[f(X)|X_{Q_A \cup S_j}] - \mathbb{E}[f(X)|X_{Q_A}] \big\|_{L^2(\mathbb{P})}.$$

It thus suffices to show that $\big\| \mathbb{E}[f(X)|X_{Q_A \cup S_j}] - \mathbb{E}[f(X)|X_{Q_A}] \big\|_{L^2(\mathbb{P})} \leq \|f(X)\|_{L^2(\mathbb{P})} = \|f\|_{L^2(P_X)}$. First, by the tower property we have $\mathbb{E}[f(X)|X_{Q_A}] = \mathbb{E}\big[\mathbb{E}[f(X)|X_{Q_A \cup S_j}]|X_{Q_A}\big]$. We then have

$$\big\| \mathbb{E}[f(X)|X_{Q_A \cup S_j}] - \mathbb{E}\big[\mathbb{E}[f(X)|X_{Q_A \cup S_j}]|X_{Q_A}\big] \big\|_{L^2(\mathbb{P})}^2 \leq Var\big(\mathbb{E}[f(X)|X_{Q_A \cup S_j}]\big) \leq Var\big(f(X)\big) \leq \|f(X)\|_{L^2(\mathbb{P})}^2,$$

where the first and second inequality are a consequence of the law of total variance and last follows from the definition of the variance. This proves part (ii).

Part (iii) follows directly from definitions.

Part (iv) can be shown by viewing $\mathbb{E}[f(X)|X_{Q_A}]$ as an orthogonal projection. First, suppose that $f_0 := \mathbb{E}[f(X)] = 0$. Then, when h is efficient, $\sum_{j=1}^m \bar{\mathcal{E}}_j^{CE}[f; h, X, \mathcal{P}] = f(X)$, and

$$\begin{aligned} \|f\|_{L^2(P_X)}^2 = \|f(X)\|_{L^2(\mathbb{P})}^2 &= \sum_{j=1}^m \langle f(X), \bar{\mathcal{E}}_j^{CE}[f; h, X, \mathcal{P}] \rangle_{L^2(\mathbb{P})} \\ &= \sum_{j=1}^m \Big(\sum_{A \subseteq M \setminus \{j\}} w(A, m) \langle f(X), \mathbb{E}[f(X)|X_{Q_A \cup S_j}] - \mathbb{E}[f(X)|X_{Q_A}] \rangle_{L^2(\mathbb{P})} \Big) \\ &= \sum_{j=1}^m \Big(\sum_{A \subseteq M \setminus \{j\}} w(A, m) \big\| \mathbb{E}[f(X)|X_{Q_A \cup S_j}] - \mathbb{E}[f(X)|X_{Q_A}] \big\|_{L^2(\mathbb{P})}^2 \Big), \end{aligned}$$

where we used the fact that $f(X)-\mathbb{E}[f(X)|X_{Q_A\cup S_j}]$, $\mathbb{E}[f(X)|X_{Q_A}]$ and $\mathbb{E}[f(X)|X_{Q_A\cup S_j}]-\mathbb{E}[f(X)|X_{Q_A}]$ are mutually orthogonal in $L^2(\mathbb{P})$.

The last expression is greater than or equal to $\sum_{j=1}^m \|\bar{\mathcal{E}}_j^{CE}[f;h,X,\mathcal{P}]\|^2_{L^2(\mathbb{P})}$ because

$$\begin{aligned}
\sum_{j=1}^m \|\bar{\mathcal{E}}_j^{CE}[f;h,X,\mathcal{P}]\|^2_{L^2(\mathbb{P})} &= \sum_{j=1}^m \Big\| \sum_{A\subseteq M\setminus\{j\}} w(A,m)\big(\mathbb{E}[f(X)|X_{Q_A\cup S_j}]-\mathbb{E}[f(X)|X_{Q_A}]\big)\Big\|^2_{L^2(\mathbb{P})}\\
&\le \sum_{j=1}^m \Big(\Big(\sum_{A\subseteq M\setminus\{j\}} w(A,m)\Big)\Big(\sum_{A\subseteq M\setminus\{j\}} w(A,m)\big\|\mathbb{E}[f(X)|X_{Q_A\cup S_j}]-\mathbb{E}[f(X)|X_{Q_A}]\big\|^2\Big)\Big)\\
&= \sum_{j=1}^m \Big(\sum_{A\subseteq M\setminus\{j\}} w(A,m)\big\|\mathbb{E}[f(X)|X_{Q_A\cup S_j}]-\mathbb{E}[f(X)|X_{Q_A}]\big\|^2\Big);
\end{aligned}$$

where we used Cauchy-Schwarz along with the non-negativity of coefficients $w(A,m)$, and $\sum_{A\subseteq M\setminus\{j\}} w(A,m) = 1$ which follows from the efficiency property. This proves (iv) for the case $f_0=0$. The case $f_0\neq 0$ is proven by replacing f above with $f-f_0$ and then using (iii) and the inequality $Var(f(X)) \le \mathbb{E}[|f(X)|^2]$.

□

C.2 Proof of Proposition 3.2

Proof. Take any $A\subseteq M$. If $f=f_*$ $\tilde{P}_{X,\mathcal{P}}$-a.s., then $f=f_*$ $P_{X_{Q_A}}\otimes P_{X_{-Q_A}}$-a.s., where $Q_A=\cup_{r\in A}S_r$. Hence

$$v^{ME,\mathcal{P}}(A\cup\{j\};X,f) = v^{ME}(Q_A\cup S_j;X,f) = v^{ME}(Q_A\cup S_j;X,f_*) = v^{ME,\mathcal{P}}(A\cup\{j\};X,f_*) \ \mathbb{P}\text{-a.s.},$$

which implies that $v^{ME,\mathcal{P}}$ is a well-defined map on $L^2(\tilde{P}_{X,\mathcal{P}})$.

Take any $f\in L^2(\tilde{P}_{X,\mathcal{P}})$. Then, for any $A\subseteq M$ we have

$$\begin{aligned}
\|v^{ME,\mathcal{P}}(A;X,f)\|^2_{L^2(\mathbb{P})} &= \mathbb{E}\big[(v^{ME}(Q_A;X;f))^2\big]\\
&= \int\Big(\int f(x_{Q_A},x_{-Q_A})P_{X_{-Q_A}}(dx_{-Q_A})\Big)^2 P_{X_{Q_A}}(dx_{Q_A}))\\
&\le \int f^2(x_{Q_A},x_{-Q_A})[P_{X_{Q_A}}\otimes P_{X_{-Q_A}}](dx_{Q_A},dx_{-Q_A}),
\end{aligned}$$

where on the last line we used Cauchy-Schwarz and the Fubini's theorem. This proves (i).

Clearly, by (i) the operator $\bar{\mathcal{E}}^{ME}[\cdot,h,X,\mathcal{P}]$ is well-defined on $L^2(\tilde{P}_{X,\mathcal{P}})$. Furthermore, for $j\in M$, we have

$$\begin{aligned}
\|\bar{\mathcal{E}}_j^{ME}[f;h,X,\mathcal{P}]\|_{L^2(\mathbb{P})} &\le \sum_{A\subseteq M\setminus\{j\}} w(A,m)\cdot\|v^{ME,\mathcal{P}}(A\cup\{j\};X,f)-v^{ME,\mathcal{P}}(A;X,f)\|_{L^2(\mathbb{P})}\\
&\le \Big(\sum_{A\subseteq M\setminus\{j\}} w^2(A,m)\Big)^{\frac{1}{2}}\Big(\sum_{A\subseteq M\setminus\{j\}}\|v^{ME,\mathcal{P}}(A\cup\{j\};X,f)-v^{ME,\mathcal{P}}(A;X,f)\|^2_{L^2(\mathbb{P})}\Big)^{\frac{1}{2}}\\
&= \Big(\sum_{A\subseteq M\setminus\{j\}} w^2(A,m)\Big)^{\frac{1}{2}}\Big(2\sum_{A\subseteq M}\|f\|^2_{L^2(P_{X_{Q_A}}\otimes P_{X_{-Q_A}})}\Big)^{\frac{1}{2}}\\
&= 2^{\frac{m+1}{2}}\Big(\sum_{A\subseteq M\setminus\{j\}} w^2(A,m)\Big)^{\frac{1}{2}}\cdot\|f\|_{L^2(\tilde{P}_{X,\mathcal{P}})},
\end{aligned}$$

where we used (3.7), which establishes (ii).

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

$$v^{ME,\mathcal{P}}(A\cup\{j\};X,f_*)-v^{ME,\mathcal{P}}(A;X,f_*) = v^{ME}(Q_A\cup\{S_j\};X,f_*)-v^{ME}(Q_A;X,f_*) = 0 \quad \mathbb{P}\text{-a.s.},$$

and hence by (3.1) we have $h[N,v^{ME,\mathcal{P}}(\cdot;h,X,f_*)]=0\in L^2(\mathbb{P})$. Since $f=f_*$ $\tilde{P}_{X,\mathcal{P}}$-a.s., using the fact that $\bar{\mathcal{E}}^{ME}[\cdot;h,X,\mathcal{P}]$ is well-defined, we conclude that $\bar{\mathcal{E}}^{ME}[f;h,X,\mathcal{P}]=0\in L^2(\mathbb{P})$, which establishes (iii). □

C.3 Proof of Lemma 3.1

Proof. As one clearly has $P_X \ll \tilde{P}_{X,\mathcal{P}}$, there is an obvious well-defined linear map $\tilde{I} : L^2(\tilde{P}_{X,\mathcal{P}}) \to L^2(P_X)$ sending the $L^2(\tilde{P}_{X,\mathcal{P}})$-class to its corresponding $L^2(P_X)$-class. The image of this linear map is the subspace $H_{X,\mathcal{P}}$ of $L^2(P_X)$, as introduced in (3.9). The image, $H_{X,\mathcal{P}}$, can be identified with the domain of $\tilde{I}$ as a vector space provided that the map $\tilde{I}$ is injective–which amounts to $\tilde{P}_{X,\mathcal{P}} \ll P_X$. Therefore, given the assumption $\tilde{P}_{X,\mathcal{P}} \ll P_X$, the subspace $H_{X,\mathcal{P}}$ of $L^2(P_X)$ is isomorphic to $L^2(\tilde{P}_{X,\mathcal{P}})$ if the latter is equipped with the same L^2 norm, thus part (i). Once $L^2(\tilde{P}_{X,\mathcal{P}})$ is identified with $H_{X,\mathcal{P}}$ as a vector space, any well-defined linear map on $L^2(\tilde{P}_{X,\mathcal{P}})$ amounts to a well-defined linear map on $H_{X,\mathcal{P}}$ in an obvious way. Parts (ii) and (iii) now follow because, according to Proposition 3.2, assignments $f \mapsto v^{ME,\mathcal{P}}(A; X, f)$ and $f \mapsto \bar{\mathcal{E}}^{ME}[f; h, X, \mathcal{P}]$ descend to well-defined operators on $L^2(\tilde{P}_{X,\mathcal{P}})$. □

C.4 Proof of Lemma 3.2

Proof. Since $\tilde{P}_{X,\mathcal{P}} \ll P_X$, for each $A \subseteq M$ we have $P_{X_{Q_A}} \otimes P_{X_{-Q_A}} \ll P_X$ and hence, by Corollary D.1, the Radon-Nikodym derivative exists and satisfies $0 \leq r_{Q_A} := \frac{dP_{X_{Q_A}} \otimes P_{X_{-Q_A}}}{dP_X} \in L^1(P_X)$. Then for any $B \in \mathcal{B}(\mathbb{R}^n)$ we have

$$\int_B r^{(\mathcal{P})}(x)\, P_X(dx) = \tilde{P}_{X,\mathcal{P}}(B) = \frac{1}{2^m} \sum_{A \subseteq M} P_{X_{Q_A}} \otimes P_{X_{-Q_A}}(B) = \frac{1}{2^m} \sum_{A \subseteq M} \int_B r_{Q_A}(x)\, P_X(dx).$$

Since $B \in \mathcal{B}(\mathbb{R}^n)$ is arbitrary, we conclude that $r^{(\mathcal{P})} = \frac{1}{2^m} \sum_{A \subseteq M} r_{Q_A}$, and by Corollary D.1, we have $\|r_S\|_{L^1(P_X)} = 1$. This proves (i).

To prove (ii), first, suppose $r^{(\mathcal{P})} \in L^\infty(P_X)$. Then, by construction, $H_{X,\mathcal{P}}$ is a subset of $L^2(P_X)$. Thus, to show that $H_{X,\mathcal{P}} = L^2(P_X)$, it suffices to show that an arbitrary function f belonging to $L^2(P_X)$ has finite $L^2(\tilde{P}_{X,\mathcal{P}})$-norm.

For any $k > 0$ we have

$$\begin{aligned}\int 1_{\{|f| \leq k\}} f^2(x)\, \tilde{P}_{X,\mathcal{P}}(dx) &= \int 1_{\{|f| \leq k\}} f^2(x) \cdot r^{(\mathcal{P})}(x)\, P_X(dx) \\ &\leq \|r^{(\mathcal{P})}\|_{L^\infty(P_X)} \int f^2(x) P_X(dx) < \infty.\end{aligned}$$

As $k \to \infty$, using the monotone convergence theorem, we obtain $f \in L^2(\tilde{P}_{X,\mathcal{P}})$. Thus, $L^2(P_X) \subseteq L^2(\tilde{P}_{X,\mathcal{P}})$. This proves that $H_{X,\mathcal{P}} = L^2(P_X)$.

Next, suppose that $H_{X,\mathcal{P}} = L^2(P_X)$. Then for $f \in L^2(P_X)$ we have

$$\infty > \int f^2(x)\, \tilde{P}_{X,\mathcal{P}}(dx) = \int f^2(x) \cdot r^{(\mathcal{P})}(x)\, P_X(dx).$$

Thus, $f \in L^2(P_X)$ implies $f \cdot (r^{(\mathcal{P})})^{1/2} \in L^2(P_X)$.

Take $g \in L^1(P_X)$. Then $|g|^{1/2} \in L^2(P_X)$, and hence $(|g| \cdot r^{(\mathcal{P})})^{1/2} \in L^2(P_X)$. Thus, for every $g \in L^1(P_X)$, we have $g \cdot r^{(\mathcal{P})} \in L^1(P_X)$. This imples that the linear functional T on $L^1(P_X)$ defined by $T(g) := \int g \cdot r^{(\mathcal{P})} P_X(dx)$ is well-defined.

Next, for each integer $k \geq 1$, define a linear functional T_k on $L^1(P_X)$ by

$$T_k(g) = \int \mathbb{1}_{\{r^{(\mathcal{P})} \leq k\}}\, r^{(\mathcal{P})} \cdot g(x)\, P_X(dx).$$

By construction, T_k is a bounded functional on $L^1(P_X)$ for each $k \in \mathbb{N}$. Thus, for any $g \in L^1(P_X)$, using the dominated convergence theorem, we have $T_k(g) \to T(g)$ as $k \to \infty$. Then, by the uniform bounded principle [53, p.269], we conclude that T must be bounded on $L^1(P_X)$. Hence, by the Reisz-Fréchet representation theorem [53, p.313], we obtain $r^{(\mathcal{P})} \in L^\infty(P_X)$. This proves (ii).

□

C.5 Proof of Proposition 3.3

Proof. Let $f \in H_{X,\mathcal{P}}$ and pick $A \subseteq M$. Then we have

$$
\begin{aligned}
&\mathbb{E}\left[(v^{CE,\mathcal{P}}(A;X,f) - v^{ME,\mathcal{P}}(A;X,f))^2\right] = \mathbb{E}_{x_{Q_A}\sim P_{X_{Q_A}}}\left[(\mathbb{E}[f(x_{Q_A},X_{-Q_A})|X_{Q_A}=x_{Q_A}] - \mathbb{E}[f(x_{Q_A},X_{-Q_A})])^2\right]\\
&= \int \Big(\int f(x_{Q_A},x_{-Q_A})P_{X_{-Q_A}|X_{Q_A}=x_{Q_A}}(dx_{-Q_A}) - \int f(x_{Q_A},x_{-Q_A})P_{X_{-Q_A}}(dx_{-Q_A})\Big)^2 P_{X_{Q_A}}(dx_{Q_A})\\
&\le \|(r_{Q_A}-1)\cdot f\|^2_{L^2(P_X)}\\
&\le \|r_{Q_A}-1\|^2_{L^\infty(P_X)}\|f\|^2_{L^2(P_X)}.
\end{aligned}
$$

The first inequality is due to Lemma D.1, since (AC) holds. The second inequality is true since (PB) holds. This proves (i). To prove (ii), we write

$$
\begin{aligned}
&\|\mathcal{I}_j(f;\{r_{Q_A}\}_{A\subseteq M},h)\|_{L^2(\mathbb{P})} = \|\bar{\mathcal{E}}_j^{CE}[f;h,\mathcal{P}] - \bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})}\\
&= \|\sum_{A\subseteq M\setminus\{j\}} w(A,m)\big(v^{CE,\mathcal{P}}(A\cup\{j\}) - v^{CE,\mathcal{P}}(A)\big) - \sum_{A\subseteq M\setminus\{j\}} w(A,m)\big(v^{ME,\mathcal{P}}(A\cup\{j\}) - v^{ME,\mathcal{P}}(A)\big)\|_{L^2(\mathbb{P})}\\
&\le \sum_{A\subseteq M\setminus\{j\}} w(A,m)\left(\|v^{CE,\mathcal{P}}(A\cup\{j\}) - v^{ME,\mathcal{P}}(A\cup\{j\})\|_{L^2(\mathbb{P})} + \|v^{CE,\mathcal{P}}(A) - v^{ME,\mathcal{P}}(A)\|_{L^2(\mathbb{P})}\right)
\end{aligned}
$$

and then we apply part (i) to obtain the result. □

C.6 Proof of Proposition 3.4

Proof. Let $f \in H_{X,\mathcal{P}}$ and pick $A \subseteq M$. To prove (i) we write

$$
\|v^{ME,\mathcal{P}}(A;X,f)\|_{L^2(\mathbb{P})} \le \|v^{ME,\mathcal{P}}(A;X,f) - v^{CE,\mathcal{P}}(A;X,f)\|_{L^2(\mathbb{P})} + \|v^{CE,\mathcal{P}}(A;X,f)\|_{L^2(\mathbb{P})}
$$

and then we apply Proposition 3.3(i) and Proposition 3.1(i) to obtain the result.

To prove (ii) we write

$$
\|\bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})} \le \|\bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}] - \bar{\mathcal{E}}_j^{CE}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})} + \|\bar{\mathcal{E}}_j^{CE}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})}
$$

and then we apply Proposition 3.3(ii) and Proposition 3.1(ii) to obtain the result.

Finally, to prove (iii) we write

$$
\begin{aligned}
\|\bar{\mathcal{E}}^{ME}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})^m} &\le \|\bar{\mathcal{E}}^{ME}[f;h,\mathcal{P}] - \bar{\mathcal{E}}^{CE}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})^m} + \|\bar{\mathcal{E}}^{CE}[f;h,\mathcal{P}]\|_{L^2(\mathbb{P})^m}\\
&\le \big(\sum_{j=1}^m \|\bar{\mathcal{E}}_j^{ME}[f;h,\mathcal{P}] - \bar{\mathcal{E}}_j^{CE}[f;h,\mathcal{P}]\|^2_{L^2(\mathbb{P})}\big)^{1/2} + \|f\|_{L^2(P_X)}
\end{aligned}
$$

where we applied Proposition 3.1(iii) to obtain the second inequality and the result follows by applying Proposition 3.3(ii) to each of the m terms in the summation. □

D 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

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

Definition D.1 (**push-forward).** Let $\mathbb{P}$ be a probability measure on a measurable space $(\Omega, \mathcal{F})$. Let $X \in \mathbb{R}^n$ be a random vector defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The push-forward probability distribution of $\mathbb{P}$ by X is defined by

$$P_X(A) := \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \in A\}\big), \quad A \in \mathcal{B}(\mathbb{R}^n).$$

Definition D.2 (**absolute continuity).** *Let μ, ν be measures on a measurable space $(\Omega, \mathcal{F})$. μ is said to be absolutely continuous with respect to ν, denoted as $\mu \ll \nu$, if $\nu(A) = 0$ implies $\mu(A) = 0$ for $A \in \mathcal{F}$.*

Theorem D.1 (**Radon-Nikodym derivative).** *Suppose that μ, ν are two σ-finite measures defined on a measurable space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$, then there exists an $\mathcal{F}$-measurable function $r : \Omega \to [0, \infty)$, written as $r = \frac{d\mu}{d\nu}$, such that for any measurable set $A \in \mathcal{F}$, $\mu(A) = \int_A r(x)\, \nu(dx)$.*

Proof. See Royden and Fitzpatrick [53]. □

Corollary D.1. *Suppose that μ, ν are two probability measures defined on a measurable space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$, then the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ belongs to $L^1(\Omega, \mathcal{F}, \nu)$ and is of norm 1.*

Lemma D.1. *Let $X = (X_1, \dots, X_n)$, $Z = (Z_1, \dots, Z_m)$ 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*

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

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 (D.1). □

Definition D.3 (**Wasserstein).** *The Wasserstein distance W_1 on $\mathscr{P}_1(\mathbb{R}^k)$ is given by [35]*

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

Lemma D.2 (**Wasserstein bound).** *Let $\mu, \nu \in \mathscr{P}_1(\mathbb{R}^k)$. Suppose $\mu \ll \nu$. Then*

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

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

$$W_1(\mu, \nu) = \sup \Big\{ \int (\psi(x) - \psi(0))(r - 1)\, \nu(dx), \ \psi \in Lip_1(\mathbb{R}^k) \Big\}. \tag{D.3}$$

Since $\psi \in Lip_1(\mathbb{R}^k)$, $|\psi(x) - \psi(0)| \leq |x|$, which implies the inequality on the right-hand side of (D.2). Next, by taking $\psi(x) = \pm|x|$, and using it in (D.3), we obtain the left-hand side of (D.2). □

E Maximal information coefficient

E.1 MIC$_e$ statistic

Definition E.1 (Reshef et al. [50]). *Let D_n be a dataset drawn from (X, Y), with $|D_n| = n$. Let $B(n)$ be a tuning function that tends to ∞ as $n \to \infty$. Then*

$$\mathrm{MIC}_e(D_n; B(n)) := \max_{k\ell < B(n)} \left\{ \frac{\mathbb{1}_{\{k<l\}} I^{[*]}(D_n, k, [\ell]) + \mathbb{1}_{\{k \geq l\}} I^{[*]}(D_n, [k], \ell)}{\log(\min\{k, \ell\})} \right\},$$

where $I^{[*]}(D_n, k, [\ell]) = \max_{G \in G(k,[l])} I(X,Y)|_G$ and $G(k, [l])$ is the set of k-by-l grids whose y-axis partition is an equipartition of size l.

Reshef et al. [50, Corrollary 26] establishes that MIC$_e$ is a consistent estimator of MIC$_*$ provided that $\omega(1) < B(n) \leq O(n^{1-\epsilon})$ for some $\epsilon \in (0, 1)$. Furthermore, Reshef et al. [50, Theorem 28] shows that MIC$_e$ can be computed in time $O(n + n^{5(1-\epsilon)/2})$ when $B(n) = O(n^{1-\varepsilon})$, which in turn implies the following.

Corollary E.1 (Reshef et al. [50]). MIC$_*$ *can be estimated consistently in linear time.*

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] M. J. Albizuri, J. M. Zarzuelo, On coalitional semivalues. *Games and Economic Behavior* 49, 221–243.

[4] R. Amer, F. Carreras, J. M. Giménez (1995). The modified Banzhaf value for games with coalition structure: an axiomatic characterization. *Mathematical Social Sciences* 43, 45–54 (1995).

[5] J. M. Alonso–Meijide, M. G. Fiestras–Janeiro, Modification of the Banzhaf value for games with a coalition structure. *Annals of Operations Research 109*, 213–227, (2002).

[6] 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.

[7] R. J. Aumann, J. Dréze Cooperative games with coalition structure. *Int. J. Game Theory*, 3, 217-237 (1974).

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

[9] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, (1985).

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

[11] B. Casas-Méndez, I. Garćıa–Jurado, A. van den Nouweland, Vázquez–Brage An extension of the τ-value to games with coalition structures. *European Journal of Operational Research* 148, 494–513, (2003).

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

[13] 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).*

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

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Ed., John Wiley & Sons, Hoboken, NJ (2006).

[16] 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.

[17] P. Dubey, A. Neyman, R. J. Weber, Value theory without efficiency. *Math. Oper. Res.* 6, 122–128, (1981).

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

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

[20] K. Filom, A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, On marginal feature attributions of tree-based models, *Foundations of Data Science, AIMS*, (to appear 2024).

[21] 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).

[22] 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).

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

[24] A. Gretton, O. Bousquet, A. Smola, and Bernhard Schölkopf Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, p. 63–77. Springer, (2005).

[25] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, (2012).

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

[27] K. Hamidieh. Superconductivty Data. UCI Machine Learning Repository, 2018. DOI: 10.24432/C53P47.

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

[29] R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, (2013).

[30] R. Heller, Y. Heller, S. Kaufman, B. Brill, and M. Gorfine. Consistent distribution-free$k$-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29):1–54, (2016).

[31] 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).

[32] 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.

[33] 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).

[34] 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*.

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

[36] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69, (2004).

[37] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Interpretable & Explorable Approximations of Black Box Models. *arXiv e-prints*, page arXiv:1707.01154, July 2017.

[38] D. Lopez-Paz, P. Hennig, and B. Schölkopf, The randomized dependence coefficient. In *Advances in Neural Information Processing Systems*, p. 1–9, (2013).

[39] S. Lorenzo-Freire, New characterizations of the Owen and Banzhaf–Owen values using the intracoalitional balanced contributions property, *TOP* 25, 579–600 (2017).

[40] S. M. Lundberg, G. G. Erion and S.-I. Lee, Consistent individualized feature attribution for tree ensembles, *arXiv preprint arxiv:1802.03888*, (2019).

[41] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, *31st Conference on Neural Information Processing Systems*, (2017).

[42] 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).

[43] 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).

[44] 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:2102.10878v6*, (2024).

[45] A. Miroshnikov, K. Kotsiopoulos, K. Filom and A. Ravi Kannan, On the stability of single-feature game-theoretic explanations of machine learning models. *GitHub preprint, (2024)*, https://github.com/alexey-miroshnikov/Stability-indiv-explanations-paper/blob/main/paper/explanation_stability_individual_main.pdf.

[46] 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).

[47] G. Owen, Modification of the Banzhaf-Coleman index for games with apriory unions. *In: Power, Voting and Voting Power (M.J. Holler, ed.), Physica-Verlag, 232-238. and Game Theory (R. Henn and O. Moeschlin, eds.)*, Springer, 76–88 (1977).

[48] L. Paninski, Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, (2003).

[49] D. N. Reshef, Y.A. Reshef, H. K. Finucane, R. S. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, (2011).

[50] 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).

[51] 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).

[52] 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.

[53] H. L. Royden, P. M. Fitzpatrick, *Real analysis*. Boston: Prentice Hall, 4th ed. (2010).

[54] A. Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3):441–451, (1959).

[55] A. Saabas. treeinterpreter python package `https://github.com/andosa/treeinterpreter`, 2019.

[56] L. S. Shapley, A value for n-person games, *Annals of Mathematics Studies*, No. 28, 307-317 (1953).

[57] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41, 3, 647-665, (2014).

[58] M. Sundararajan, A. Najmi, The Many Shapley Values for Model Explanation, *International conference on machine learning*, pages 9269–9278, PMLR, (2020).

[59] G. J. Szekely, M. L. Rizzo, N. Bakirov, Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, (2007).

[60] G. J. Szekely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236-1265, (2009).

[61] 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.

[62] S. H. Tijs (1981), Bounds for the core and the *tau*-value. *In: Game Theory and Mathematical Economics (O. Moeschlin and D. Pallaschke, eds.)*, North–Holland, 123–132.

[63] 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).

[64] J. J. Vidal-Puga, The Harsanyi paradox and the right to talk in bargaining among coalitions. *Mathematical Social Sciences* 64, 214-224, (2012).

[65] J. Wang, J. Wiens, S. Lundberg Shapley Flow: A Graph-based Approach to Interpreting Model Predictions *arXiv preprint arXiv:2010.14592*, (2020).

[66] I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2009. DOI: 10.24432/C55S3H.

[67] X. Zenga, Y. Xiaa, and H. Tong, Jackknife approach to the estimation of mutual information, *PNAS* , 115-40, (2019).