SUPPLEMENTARY MATERIALS: Stability of game-theoretic feature explanations for machine learning models

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

SM1. Pedagogical example on instability of marginal explanations. This section contains an example that illustrates the theoretical aspects of instability discussed in the paper.

The results of §3 show that marginal explanations viewed as linear operators may not be well-defined or continuous in $L^2(P_X)$. Demonstrating the instability 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 explanations by training a small collection of models on perturbed datasets, and then comparing the differences between the resulting explanations as well as between the predictions of those models.

In what follows, we consider the following data generating model. 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

$$\begin{aligned} &Z \sim Unif(-1,1)\\ &X_1 = Z + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0,\delta),\\ &X_2 = \sqrt{2}\sin(Z(\pi/4)) + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0,\delta),\\ &X_3 \sim Unif\big([-1,-0.5]\cup[0.5,1]\big). \end{aligned} \tag{SM1.1}$$

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

$$Y = f_*(X_1, X_2, X_3) = 3X_2X_3. \tag{SM1.2}$$

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

By design, the dependencies in predictors allow for the existence of many models from $L^2(P_X)$ that approximate the response variable well but have different representations. In what follows, we demonstrate that the generated explanations differ in such cases where different models with distinct representations approximate the data well.

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

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

distribution to the training sets, have widely different explanations. It also illustrates how grouping features based on dependencies rectifies the explanation instabilities. The details of the experiment are provided below.

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 (SM1.1), 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 (SM1.1) with noise $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \delta)$, and Y_δ is constructed using the response model (SM1.2). 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[1]: 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 (SM1.2). The mixture dataset D is used later for computing predictive performance of each $f \in \{f_*, f_1, \dots, f_5\}$ on D and the estimation of corresponding marginal explanations for (X, f) across D.

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 $\|f_i\|_{L^2(P_X)}$ of the models recorded in Table SM1, which illustrates 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 of the distances are given by

$$\|f_k - f_*\|_{L^2(P_X)} \approx (0.069, 0.062, 0.056, 0.071, 0.064), \quad \|f_*\|_{L^2(P_X)} \approx 1.37, \tag{SM1.3}$$

and also recorded in Table SM1. Thus, the predictions of the trained models on the mixture dataset are close in an L^2-sense to those of f_*. In particular, this implies that $\{f_k\}_{k=1}^5$ live 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.

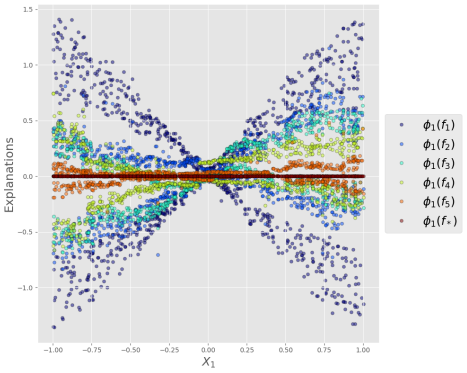
We next pick $m = 1000$ samples at random from the mixture dataset, to construct the dataset $D_X^{(e)}$ of predictor observations used for explanations. We also subsample the predictors from the mixture set and obtain a background dataset $\bar{D}_X$ with 1000 samples. The background dataset is used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

$$\hat{v}^{ME}(S; x, f, \bar{D}_X) := \frac{1}{|\bar{D}_X|} \sum_{\tilde{x} \in \bar{D}_X} f(x_S, \tilde{x}_{-S}) \approx \mathbb{E}[f(x_S, X_{-S})],$$

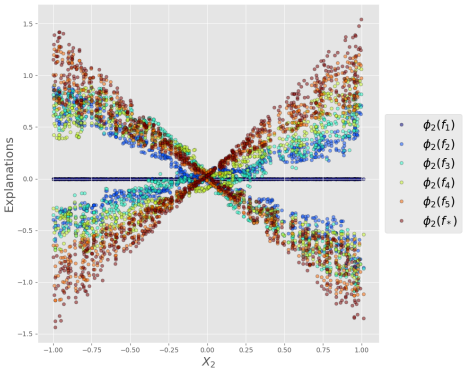
where $x \in D_X^{(e)}$ is an observation and $f \in \{f_*, f_1, \dots f_5\}$.

We then evaluate the empirical marginal explanations $\varphi_i[N, \hat{v}^{ME}](x)$ 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 [1], which computes empirical marginal Shapley values for tree-based models such as XGBoost.

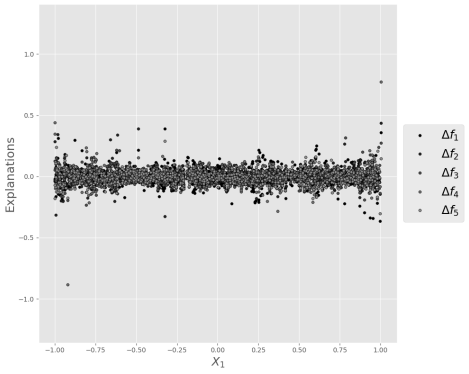
[1]The code is available at https://github.com/alexey-miroshnikov/Stability-indiv-explanations-paper.



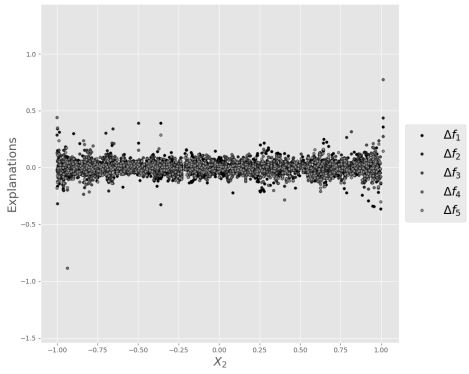
(a) Explanations φ_1 vs X_1.



(b) Explanations φ_2 vs X_2.

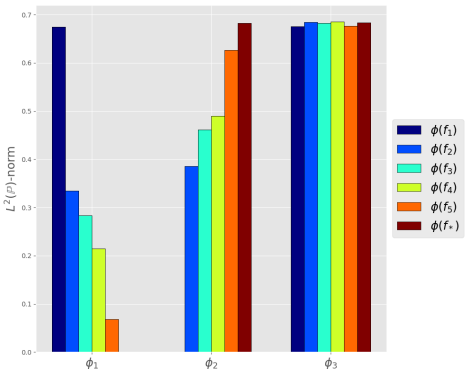


(c) Differences of predictions vs X_1.

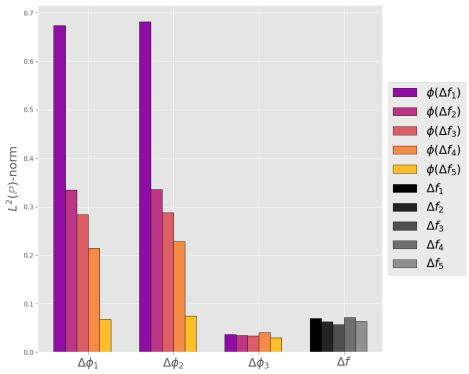


(d) Differences of predictions vs X_2.

Figure SM1: Individual marginal explanations and response.



(a) Explanation norms.



(b) Global explanations of Δf_i.

Figure SM2: Individual explanation norms.

	$\|\cdot\|$	β_1	β_2	β_3	$\|\beta\|$
f_1	1.370	0.674	0.000	0.675	0.954
f_2	1.375	0.334	0.385	0.685	0.854
f_3	1.374	0.285	0.461	0.682	0.871
f_4	1.375	0.214	0.228	0.040	0.315
f_5	1.374	0.068	0.627	0.677	0.925
f_*	1.380	0.000	0.682	0.682	0.965
$f_1 - f_*$	0.069	0.674	0.682	0.036	0.960
$f_2 - f_*$	0.062	0.334	0.336	0.035	0.475
$f_3 - f_*$	0.056	0.284	0.288	0.033	0.406
$f_4 - f_*$	0.071	0.214	0.228	0.041	0.315
$f_5 - f_*$	0.064	0.067	0.075	0.029	0.104

Table SM1: Global marginal Shapley attributions.

Figures SM1a-SM1b depict the scatterplots of explanations for each model across the dataset $D_X^{(e)}$, where we see that explanations differ substantially, indicating that the trained models have different functional representations. In particular, f_1 treats both predictors X_1, X_2 similarly due to the strong dependence between them, while f_5 treats the first predictor as a dummy variable which is similar to the model f_*.

To quantify the global attribution of each predictor, we estimate the L^2-norms of the marginal Shapley values for each model, $\beta_i(f_k, \hat{v}^{ME}) := \|\varphi_i(X; f_k, \hat{v}^{ME})\|_{L^2(\mathbb{P})}$, $i \in N$, which are depicted in Figure SM2a and recorded in Table SM1. These values also demonstrate that the features X_1, X_2 are utilized differently across the models.

Recall that by Corollary 3.5(ii) (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 $|\beta(f_1 - f_2, v^{CE})|/\|f_1 - f_2\|_{L^2(P_X)} \leq 1$, where $\beta := (\beta_1, \beta_2, \beta_3)$. This bound ensures that the total distance $|\beta(f_1 - f_2, v^{CE})|$ 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 the vector $\beta(f_1 - f_2, v^{CE})$. Meanwhile, in theory, in the presence of dependencies, the bound for the marginal explanations may in general be infinite or significantly larger than one, which depends on the relationship between P_X and $\tilde{P}_X$.

To understand 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 $\beta_i(f_i - f_*, \hat{v}^{ME}) = \|\varphi_i(X; f_i - f_*, \hat{v}^{ME})\|_{L^2(\mathbb{P})}$, and then compare with those of the model. Figure SM2b, where $f_k - f_*$ is denoted as Δf_k, showcases the comparison between the distances of the individual feature explanations and the model distances, again for each trained model.

We contrast the unit operator bound in (3.5) 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 Figure SM2b. 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; see Figure SM2b. We note that the total distance between explanations is significant and, in particular, for the model f_1 it constitutes about 60% of the trained models' norm; see Table SM1.

SM2. On the relationship between probability measures P_X and $\tilde{P}_X$. The comparison of probability measures P_X and $\tilde{P}_X := \frac{1}{2^n}\sum_{S\subseteq N} P_{X_S}\otimes P_{X_{-S}}$ lies at the heart of the analysis of conditional and marginal explanations carried out in this paper. Recall that the former is the joint probability distribution of predictors $X = (X_1, \ldots, X_n)$ while the latter probability measure on $\mathbb{R}^n$ emerged naturally in our investigation of marginal explanations.

Proposition SM2.1. *The following three statements are equivalent.*

(a) The predictors are independent.

(b) $P_{X_S}\otimes P_{X_{-S}}$ coincides with P_X for every $S \subseteq N$.

(c) $\tilde{P}_X$ coincides with P_X.

Proof. Obviously $(a) \implies (b) \implies (c)$. It remains to show that $(c) \implies (a)$. We prove this by induction on n. First, we claim that if $\tilde{P}_X = P_X$ where $X = (X_1, \ldots, X_n)$, then any $n-1$ of these random variables are independent. By symmetry, it suffices to show that $X_1, \ldots, X_{n-1}$ are independent. Let $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ denote the projection onto the first $n-1$ coordinates. Then $\pi_* P_X = P_{X'}$ where $X' := (X_1, \ldots, X_{n-1})$. Also the pushforward of $\tilde{P}_X = \frac{1}{2^n}\sum_{S\subseteq N} P_{X_S}\otimes P_{X_{N\setminus S}}$ by π is equal to $\tilde{P}_{X'} = \frac{1}{2^{n-1}}\sum_{S\subseteq N'} P_{X_S}\otimes P_{X_{N'\setminus S}}$ where $N' := \{1, \ldots, n-1\}$. This is due to the fact that for every $S \subseteq N'$, $P_{X_S}\otimes P_{X_{N'\setminus S}}$ can be realized as the pushforward of two terms in $\tilde{P}_X$: $P_{X_S}\otimes P_{X_{N\setminus S}}$ and $P_{X_{S\cup\{n\}}}\otimes P_{X_{N\setminus(S\cup\{n\})}}$. Consequently, applying π_* to $\tilde{P}_X = P_X$ yields $\tilde{P}_{X'} = P_{X'}$, and thus by the induction hypothesis, the independence of $X_1, \ldots, X_{n-1}$. Now since any $n-1$ of the random variables $X_1, \ldots, X_n$ are independent, for any non-empty and proper subset S of N we have $P_{X_S}\otimes P_{X_{N\setminus S}} = P_{X_1}\otimes\cdots\otimes P_{X_n}$. When $S = \varnothing$ or N, the measure $P_{X_S}\otimes P_{X_{N\setminus S}}$ coincides with P_X. Therefore, $\tilde{P}_X = P_X$ amounts to

$$\frac{1}{2^n}\big((2^n-2)P_{X_1}\otimes\cdots\otimes P_{X_n} + 2P_X\big) = P_X$$

which results in $P_{X_1}\otimes\cdots\otimes P_{X_n} = P_X$, i.e. random variables $X_1, \ldots, X_n$ are independent. ∎

Next, we move from equality $\tilde{P}_X = P_X$ to the continuity condition $\tilde{P}_X \ll P_X$. The probability measure $\tilde{P}_X$ is a convex combination of the product measures $P_{X_S}\otimes P_{X_{-S}}$. The latter is P_X when $S = \varnothing$ or N which immediately indicates that the other direction holds: $P_X \ll \tilde{P}_X$. The condition $\tilde{P}_X \ll P_X$ amounts to $P_{X_S}\otimes P_{X_{-S}} \ll P_X$ for all $S \subseteq N$. As discussed extensively in the paper, this condition appears when it comes to setting up marginal explanations as well-defined operators. The goal here is to elaborate on it through providing some examples and non-examples.[2] Especially, we elucidate this condition by relating it to the shape of the support of P_X. Recall that the support $\mathrm{supp}(\mu)$ of a Borel measure μ on

[2]Inspired by this problem, we had raised a question on MathOverflow [2].

a metric space is the set of points whose every open neighborhood has a positive measure [4]. Its complement is thus the union of all measure zero open subsets. Hence $\mathrm{supp}(\mu)$ is automatically closed; and in the case of a separable space such as $\mathbb{R}^n$, the support can be characterized as the complement of the largest open subset of measure zero.

Lemma SM2.2. *One always has* $\mathrm{supp}(P_X) \subseteq \mathrm{supp}(\tilde{P}_X)$ *and the supports coincide if* $\tilde{P}_X \ll P_X$. *Moreover, if* $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$, *then for any* $\varnothing \neq S \subset N$, *they coincide with* $\mathrm{supp}\big(P_{X_S} \otimes P_{X_{-S}}\big)$ *and* $\pi_S\big(\mathrm{supp}(P_X)\big) \times \pi_{-S}\big(\mathrm{supp}(P_X)\big)$ *where* $\pi_S : \mathbb{R}^n \to \mathbb{R}^{|S|}$ *and* $\pi_{-S} : \mathbb{R}^n \to \mathbb{R}^{n-|S|}$ *are projections onto coordinates belonging or not belonging to* S *respectively.*[3]

Proof. For any two Borel measures μ and ν on $\mathbb{R}^n$, $\mu \ll \nu$ implies $\mathrm{supp}(\mu) \subseteq \mathrm{supp}(\nu)$. Thus $\mathrm{supp}(P_X) \subseteq \mathrm{supp}(\tilde{P}_X)$ due to $P_X \ll \tilde{P}_X$; and also $\tilde{P}_X \ll P_X$ yields $\mathrm{supp}(\tilde{P}_X) \subseteq \mathrm{supp}(P_X)$, and hence $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$. Next, suppose $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$. These sets should contain $\mathrm{supp}(P_{X_S} \otimes P_{X_{-S}})$ for any S because $P_{X_S} \otimes P_{X_{-S}} \ll \tilde{P}_X$. It follows easily from the definition of a measure's support that $\mathrm{supp}(P_{X_S} \otimes P_{X_{-S}}) = \mathrm{supp}(P_{X_S}) \times \mathrm{supp}(P_{X_{-S}})$ and $\mathrm{supp}(P_{X_{\pm S}}) \supseteq \pi_{\pm S}\big(\mathrm{supp}(P_X)\big)$. Therefore:

$$\pi_S\big(\mathrm{supp}(P_X)\big) \times \pi_{-S}\big(\mathrm{supp}(P_X)\big) \subseteq \mathrm{supp}(P_{X_S} \otimes P_{X_{-S}}) \subseteq \mathrm{supp}(\tilde{P}_X) = \mathrm{supp}(P_X).$$

But clearly $\mathrm{supp}(P_X) \subseteq \pi_S\big(\mathrm{supp}(P_X)\big) \times \pi_{-S}\big(\mathrm{supp}(P_X)\big)$. Consequently, all the subsets appeared above coincide. ∎

The lemma clearly shows that $\tilde{P}_X \ll P_X$ requires the support of P_X to have a "product structure".

Corollary SM2.3. *If* $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$, *then* $\mathrm{supp}(X) = \prod_{i \in N} \pi_i\big(\mathrm{supp}(X)\big)$ *where* π_i *denotes the projection onto the* i^{th} *coordinate. In particular, this holds when* $\tilde{P}_X \ll P_X$.

Proof. Follows from fact that $\mathrm{supp}(P_X) = \pi_S\big(\mathrm{supp}(P_X)\big) \times \pi_{-S}\big(\mathrm{supp}(P_X)\big)$ for all subsets $\varnothing \neq S \subset N$ if $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$. ∎

The product structure $\mathrm{supp}(X) = \prod_{i \in N} \pi_i\big(\mathrm{supp}(X)\big)$ puts a constraint on the support: Its projections to coordinate axes must be closed[4], something which does not hold generally for an arbitrary closed subset of $\mathbb{R}^n$. In terms of the joint probability, the product structure means that the predictors take their values "heterogenously": Given numbers $a_1, \ldots, a_n$, if for every $\epsilon > 0$ there is a positive probability of X_i lying in $(a_i - \epsilon, a_i + \epsilon)$, then the probability of $(X_1, \ldots, X_n)$ belonging to any given open neighborhood of $(a_1, \ldots, a_n)$ is non-zero. In contrast, when the data lies on a "complicated" lower-dimensional submanifold of $\mathbb{R}^n$, we are in a different regime where $\tilde{P}_X \ll P_X$ fails. This last assertion is made rigorous below:

Corollary SM2.4. *If* $\mathrm{supp}(P_X) \subsetneq \mathbb{R}^n$ *is not a Cartesian product of* n *subsets of* $\mathbb{R}$, *then* $\tilde{P}_X$ *cannot be absolutely continuous with respect to* P_X. *In particular, when* $\mathrm{supp}(P_X)$ *is connected, the continuity fails unless* $\mathrm{supp}(P_X)$ *is a (possibly degenerate or unbounded or both) rectangular cube.*

[3] Following our convention, ignoring the order of coordinates, a vector $x \in \mathbb{R}^n$ may be written as (x_S, x_{-S}), and this is how $\mathrm{supp}\big(P_{X_S} \otimes P_{X_{-S}}\big) = \pi_S\big(\mathrm{supp}(X)\big) \times \pi_{-S}\big(\mathrm{supp}(X)\big)$ should be understood.

[4] Choosing arbitrary points $a_i \in \pi_i\big(\mathrm{supp}(X)\big)$, due to this product decomposition, each $\pi_i\big(\mathrm{supp}(X)\big)$ is the preimage of the closed subset $\mathrm{supp}(X)$ under the continuous map $\mathbb{R} \to \mathbb{R}^n : t \mapsto (a_1, \ldots, a_{i-1}, t, a_{i+1}, \ldots, a_n)$.

Proof. As established above, $\tilde{P}_X \ll P_X$ yields the equality $\mathrm{supp}(X) = \prod_{i\in N} \pi_i\big(\mathrm{supp}(X)\big)$, which requires all subsets appearing on the right-hand side to be closed. If the support is connected, each projection $\pi_i\big(\mathrm{supp}(X)\big)$ of it must be a connected subset of $\mathbb{R}$, i.e. an interval (closed and possibly degenerate). Therefore, $\mathrm{supp}(X)$ is a product of intervals in that case. ■

Finally, we discuss the converse implication: Can the continuity of measures be deduced from assumptions about the supports? As a matter of fact, the equality of supports $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$–which as we saw is a necessary condition for $\tilde{P}_X \ll P_X$, and implies that $\mathrm{supp}(P_X)$ has a product structure–can yield $\tilde{P}_X \ll P_X$ if the features are discrete, or admit a density function (with a small caveat, see below).

Proposition SM2.5. *The equality of supports* $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$ *implies the continuity of measures* $\tilde{P}_X \ll P_X$ *under any of the following assumptions on the predictors:*

(i) The support of each X_i *is a discrete subset of* $\mathbb{R}$.

(ii) The joint probability distribution P_X *of* $(X_1, \ldots, X_n)$ *admits a density function which is Lebesgue a.e. positive on* $\mathrm{supp}(P_X)$.

Proof. When the closed subset $\mathrm{supp}(X_i)$ is discrete, the probability of X_i belonging to a Borel subset of $\mathbb{R}$ is positive if and only if it intersects $\mathrm{supp}(X_i)$. The same is true for any random vector X_S ($S \subseteq N$) in place of X_i because $\mathrm{supp}(X_S)$ (being contained in $\prod_{i\in S} \mathrm{supp}(X_i)$) is discrete too. Pick a subset $\varnothing \neq S \subset N$. It suffices to show $P_{X_S} \otimes P_{X_{-S}} \ll P_X$; that is, $P_{X_S} \otimes P_{X_{-S}}(B) = 0$ for any Borel subset B of $\mathbb{R}^n$ with $P_X(B) = 0$. As discussed above, B does not intersect $\mathrm{supp}(P_X)$. But this subset, according to the lemma, coincides with $\mathrm{supp}(P_{X_S} \otimes P_{X_{-S}})$ because the hypothesis is that $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$. So B cannot intersect $\mathrm{supp}(P_{X_S} \otimes P_{X_{-S}})$ either. This support is discrete as well (being equal to $\mathrm{supp}(P_{X_S}) \times \mathrm{supp}(P_{X_{-S}})$). We deduce that $P_{X_S} \otimes P_{X_{-S}}(B) = 0$, as desired.

For the second part, let ρ be a density for P_X, a Borel measurable function $\rho : \mathbb{R}^n \to [0, \infty)$. Fix a subset $\varnothing \neq S \subset N$. The product measure $P_{X_S} \otimes P_{X_{-S}}$ admits a density function of form $x \mapsto \rho_S(x_S)\rho_{-S}(x_{-S})$ where $\rho_S(x_S) := \int \rho(x_S, x_{-S}) dx_{-S}$ and $\rho_{-S}(x_{-S}) := \int \rho(x_S, x_{-S}) dx_S$. When a density exists, the measure of a Borel subset is zero if and only the density vanishes at Lebesgue-almost every point of it. Therefore, to establish $P_{X_S} \otimes P_{X_{-S}} \ll P_X$, we only need to show that $P_{X_S} \otimes P_{X_{-S}}\big(\{x \in \mathbb{R}^n \mid \rho(x) = 0\}\big) = 0$, or equivalently the Lebesgue measure of $\{x \in \mathbb{R}^n \mid \rho(x) = 0, \rho_S(x_S)\rho_{-S}(x_{-S}) \neq 0\}$ is zero. This subset is contained in the union

$$\big\{x \in \mathrm{supp}(P_X) \mid \rho(x) = 0\big\} \cup \big\{x \in \mathbb{R}^n \setminus \mathrm{supp}(P_X) \mid \rho_S(x_S)\rho_{-S}(x_{-S}) \neq 0\big\}.$$

The first subset is of Lebesgue measure zero due to our assumption. Proving the same for the second one concludes the proof. As argued previously in this proof, $\mathrm{supp}(P_X)$ coincides with $\mathrm{supp}(P_{X_S} \otimes P_{X_{-S}}) = \mathrm{supp}(P_{X_S}) \times \mathrm{supp}(P_{X_{-S}})$ because of $\mathrm{supp}(P_X) = \mathrm{supp}(\tilde{P}_X)$. Hence $\{x \in \mathbb{R}^n \setminus \mathrm{supp}(P_X) \mid \rho_S(x_S)\rho_{-S}(x_{-S}) \neq 0\}$ is contained in the union

$$\big\{x \in \mathbb{R}^n \mid x_S \notin \mathrm{supp}(P_{X_S}), \rho_S(x_S) \neq 0\big\} \cup \big\{x \in \mathbb{R}^n \mid x_{-S} \notin \mathrm{supp}(P_{X_{-S}}), \rho_{-S}(x_{-S}) \neq 0\big\}.$$

They are both of Lebesgue measure zero in $\mathbb{R}^n$ since subsets $\{\rho_S \neq 0\} \setminus \mathrm{supp}(P_{X_S})$ and $\{\rho_{-S} \neq 0\} \setminus \mathrm{supp}(P_{X_{-S}})$ are of Lebesgue measure zero in the corresponding Euclidean spaces $\mathbb{R}^{|S|}$ and $\mathbb{R}^{n-|S|}$ due to the fact that ρ_S and ρ_{-S} are respectively density functions for probability measures P_{X_S} on $\mathbb{R}^{|S|}$ and $P_{X_{-S}}$ on $\mathbb{R}^{n-|S|}$. ■

Example SM2.1. We provide an example to show that the condition from the second part of theorem above on the values that the density function assumes on the support is necessary. Let $C \subset [0,1]$ be a "fat" Cantor set, i.e. a Cantor set of positive Lebesgue measure $\alpha \in (0,1)$. Let the density function of $X = (X_1, X_2)$ be $\rho := \frac{1}{1-\alpha^2} \cdot \mathbb{1}_{[0,1]^2 \setminus C^2}$. So the probability distribution P_X is continuous with respect to the Lebesgue measure, and its support is the whole square $[0,1]^2$ because C^2 is a closed and nowhere-dense subset of the square. But ρ vanishes on the subset C^2 which is of positive Lebesgue measure. We argue that $P_{X_1} \otimes P_{X_2}(C^2)$, unlike $P_X(C^2)$, is non-zero. A density function for $P_{X_1} \otimes P_{X_2}$ is $(x_1, x_2) \mapsto \tilde{\rho}(x_1)\tilde{\rho}(x_2)$ where

$$\tilde{\rho}(t) := \frac{1}{1-\alpha^2} \cdot \begin{cases} 1 & t \in [0,1] \setminus C, \\ 1-\alpha & t \in C. \end{cases}$$

This density of $P_{X_1} \otimes P_{X_2}$ is positive at every point of $[0,1]^2$ which yields $\operatorname{supp}(P_{X_1} \otimes P_{X_2}) = [0,1]^2$, and $P_{X_1} \otimes P_{X_2}(C^2) > 0$ because the two-dimensional Lebesgue measure of C^2 is positive. Consequently, continuous probability distributions P_X and $\tilde{P}_X = \frac{1}{2}\Big(P_X + P_{X_1} \otimes P_{X_2}\Big)$ have the same support $[0,1]^2$ while $\tilde{P}_X \not\ll P_X$ due to the fact that

$$P_X(C^2) = 0 < \tilde{P}_X(C^2).$$

SM3. On condition (UO). In our investigation of the marginal explanation operators, in Theorem 3.16, we set forth a condition that, if holds, causes the operators to be unbounded with respect to the $\|\cdot\|_{L^2(P_X)}$ norm even when $\tilde{P}_X \ll P_X$. Recall the (UO) (Unbounded Operator):

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

Theorem 3.16 asserts that, given predictors $X = (X_1, \ldots, X_n)$ and a game value $(N, v) \mapsto h[N, v] = (h_i[N, v])_{i \in N}$ whose coefficients satisfy a positivity condition specified therein, if (UO) is satisfied for distinct indices $i, j \in N$, then the associated maps $f \mapsto \bar{\mathcal{E}}_i^{ME}[f; h, X]$ and $f \mapsto \bar{\mathcal{E}}_j^{ME}[f; h, X]$ are unbounded when the domain is equipped with $\|\cdot\|_{L^2(P_X)}$. Here, we point out that the expression in (UO) emerges naturally when h is the Shapley value φ (whose coefficients are of course positive). With $R = A \times B$, and setting $f_R(x) := \mathbb{1}_R(x_i, x_j)$, we shall argue that

$$\text{(SM3.1)} \qquad \frac{\|\bar{\mathcal{E}}_i^{ME}[f_R; \varphi, X]\|^2_{L^2(\mathbb{P})}}{\|f_R\|^2_{L^2(P_X)}} = \frac{1}{4} \frac{[P_{X_i} \otimes P_{X_j}](R)}{P_{(X_i,X_j)}(R)} \big(P_{X_i}(A) + P_{X_j}(B)\big) + O(1)$$

as A and B vary among Borel subsets of $\mathbb{R}$ with $P_{(X_i,X_j)}(A \times B) > 0$. This will indicate that for the Shapley value, the unboundedness of marginal explanations, at least once restricted to indicator functions, results in condition (UO) from the paper—hence motivating condition (UO). To establish the equality above, we revisit the following from the proof of Theorem

3.16:

$$\begin{aligned}\bar{\mathcal{E}}_i^{ME}[f_R; h, X] &= w_{\{i,j\}}\Big(v^{ME}(\{i\}; X, f_R) - v^{ME}(\varnothing; X, f_R)\Big) \\ &\quad + w_{\{i\}}\Big(v^{ME}(\{i,j\}; X, f_R) - v^{ME}(\{j\}; X, f_R)\Big) \\ &= w_{\{i,j\}}\Big(\mathbb{1}_A(X_i)P_{X_j}(B) - P_{(X_i,X_j)}(R)\Big) \\ &\quad + w_{\{i\}}\Big(\mathbb{1}_R(X_i, X_j) - \mathbb{1}_B(X_j)P_{X_i}(A)\Big)\end{aligned}$$

where the $w_{\{i\}}$ and $w_{\{i,j\}}$ are defined it terms of the coefficients $w(S, n)\,(S \subset N)$ of the game value h as:

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

When $h = \varphi$, the coefficients are given by $w(S, n) = \frac{1}{n\binom{n-1}{|S|}}$, and:

$$\begin{aligned}w_{\{i,j\}} &= \sum_{s=0}^{n-2} \frac{1}{n\binom{n-1}{s}} \cdot \binom{n-2}{s} = \sum_{s=0}^{n-2} \frac{n-s-1}{n(n-1)} = \frac{(n-1)+\cdots+1}{n(n-1)} = \frac{1}{2}, \\ w_{\{i\}} &= \sum_{s=1}^{n-1} \frac{1}{n\binom{n-1}{s}} \cdot \binom{n-2}{s-1} = \sum_{s=1}^{n-1} \frac{s}{n(n-1)} = \frac{1+\cdots+(n-1)}{n(n-1)} = \frac{1}{2}.\end{aligned}$$

Substituting in the formula above, we have

$$\|\bar{\mathcal{E}}_i^{ME}[f_R; \varphi, X]\|^2_{L^2(\mathbb{P})} = \frac{1}{4} \cdot \mathbb{E}\Big[\big(\mathbb{1}_A(X_i)P_{X_j}(B) - P_{(X_i,X_j)}(R) + \mathbb{1}_R(X_i, X_j) - \mathbb{1}_B(X_j)P_{X_i}(A)\big)^2\Big]$$

which can be simplified as

$$\frac{1}{4}\Big(P_{X_i}(A)P_{X_j}(B)^2 + P_{X_i}(A)^2P_{X_j}(B)\Big) + P_{(X_i,X_j)}(R) \cdot (\text{a bounded term})$$

where the bounded term in parentheses is

$$\frac{1}{4}\Big(1 - P_{(X_i,X_j)}(R) + 2P_{X_j}(B) - 2P_{X_i}(A) - 2P_{X_i}(A)P_{X_j}(B)\Big) \in (-1, 1).$$

Dividing by $\|f_R\|^2_{L^2(P_X)} = P_{(X_i,X_j)}(R)$, we arrive at (SM3.1), as desired.

SM4. On H_X and the Radon-Nikodym derivative $r = \frac{d\tilde{P}_X}{dP_X}$. In Theorem 3.17 we established that if $r = \frac{d\tilde{P}_X}{dP_X}$ exists and belongs to $L^\infty(P_X)$, then $H_X = L^2(P_X)$ where

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

It turns out that the reverse is true as well. Specifically, we have the following.

Lemma SM4.1. *Suppose $\tilde{P}_X \ll P_X$ and $r := \frac{d\tilde{P}_X}{dP_X}$. The following statements are equivalent:*
(i) $r \in L^\infty(P_X)$.
(ii) $H_X = L^2(P_X)$.

Proof. First, suppose $r \in L^\infty(P_X)$. By construction, H_X is a subset of $L^2(P_X)$. Thus, to show that $H_X = L^2(P_X)$ it suffices to show that $L^2(P_X) \subseteq L^2(\tilde{P}_X)$. Pick any $f \in L^2(P_X)$. For any $k > 0$ we have

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

Then sending $k \to \infty$ and using the monotone convergence theorem we conclude that $f \in L^2(\tilde{P}_X)$. Thus, $L^2(P_X) \subseteq L^2(\tilde{P}_X)$. This proves that $H_X = L^2(P_X)$.

Next, suppose that $H_X = L^2(P_X)$. Then for every $f \in L^2(P_X)$ we have

$$\infty > \int f^2(x)\tilde{P}_X(dx) = \int f^2(x) r(x) P_X(dx).$$

Thus, for every $f \in L^2(P_X)$, we have $fr^{1/2} \in L^2(P_X)$.

Set $A_k := \{x \in \mathbb{R}^n : r(x) \geq k \ P_X\text{-a.s.}\}$ for every nonnegative integer $k \geq 0$. Suppose r is not P_X-essentially bounded. Then $P_X(A_k) > 0$ for every $k \geq 0$ and

$$f_*(x) := \left(\sum_{k=1}^{\infty} \frac{1}{k^2} 1_{A_k}(x) \frac{1}{P_X(A_k)} \right)^{1/2}$$

is well-defined. Then, by the monotone convergence theorem we have

$$\int f_*^2(x) P_X(dx) = \int \left(\sum_{k=0}^{\infty} \frac{1}{k^2} 1_{A_k}(x) \frac{1}{P_X(A_k)} \right) P_X(dx) = \sum_{k=0}^{\infty} \frac{1}{k^2} < \infty.$$

Thus, $f_* \in L^2(P_X)$. However, for every $K \geq 0$ we have

$$\int f_*^2(x) r(x) P_X(dx) \geq \int \Big(\sum_{k=0}^{K} \frac{1}{k^2} 1_{A_k}(x) \frac{1}{P_X(A_k)} \Big) r(x) P_X(dx) \geq \sum_{k=0}^{K} \frac{1}{k}.$$

Sending $K \to \infty$, we conclude that $f_* r^{1/2} \notin L^2(P_X)$, which is a contradiction. Hence r is P_X-essentially bounded. ∎

In the above proof, to construct f_*, we used help from MathOverflow [3]. As [3] points out, an alternative proof is to show that r induces a bounded linear functional on $L^1(P_X)$ using the uniform boundedness principle and then apply the Riesz representation theorem.

SM5. On boundedness of the marginal game. In Lemma 3.15, we established that if for some nonempty $S \subset N$ the Radon-Nikodym derivative $r_S = \frac{d[P_{X_S} \otimes P_{X_{-S}}]}{dP_X} \in L^\infty(P_X)$, then $v^{ME}(S; X, \cdot)$ is bounded on H_X. In this section, we show that P_X-essential boundedness of r_S is not necessary for the marginal game to be bounded on H_X.

To simplify the analysis, let $n = 2$ and suppose P_X has the density

$$p(x_1, x_2) := \mathbb{1}_{[0,1]^2} \cdot 3(x_1^2 x_2 + x_1 x_2^2)$$

which is strictly positive on $(0,1)^2$, the interior of $\text{supp} P_X = [0,1]^2$, and zero outside of it.

Clearly, the marginal measures P_{X_1} and P_{X_2} have densities

$$p_1(x_1) := \mathbb{1}_{[0,1]} \cdot 3\big(\tfrac{1}{2}x_1^2 + \tfrac{1}{3}x_1\big), \quad p_2(x_2) := \mathbb{1}_{[0,1]} \cdot 3\big(\tfrac{1}{2}x_2^2 + \tfrac{1}{3}x_2\big).$$

Hence for $x = (x_1, x_2) \in (0,1)^2$ we have

$$\frac{p_1(x_1)p_2(x_2)}{p(x_1, x_2)} = 3\frac{\big(\frac{1}{2}x_1^2 + \frac{1}{3}x_1\big)\big(\frac{1}{2}x_2^2 + \frac{1}{3}x_2\big)}{(x_1^2 x_2 + x_1 x_2^2)} = 3\frac{\big(\frac{1}{2}x_1 + \frac{1}{3}\big)\big(\frac{1}{2}x_2 + \frac{1}{3}\big)}{(x_1 + x_2)}. \tag{SM5.1}$$

Thus, for $x = (x_1, x_2) \in (0,1)^2$ we have

$$\frac{p_1(x_1)p_2(x_2)}{p(x_1, x_2)} p_2(x_2) = 9\frac{\big(\frac{1}{2}x_1 + \frac{1}{3}\big)\big(\frac{1}{2}x_2 + \frac{1}{3}\big)}{(x_1 x_2^{-1} + 1)}\big(\tfrac{1}{2}x_2 + \tfrac{1}{3}\big) \le 9 \cdot \Big(\frac{5}{6}\Big)^3 =: c_*. \tag{SM5.2}$$

and

$$\frac{p_1(x_1)p_2(x_2)}{p(x_1, x_2)} p_1(x_1) = 9\frac{\big(\frac{1}{2}x_1 + \frac{1}{3}\big)\big(\frac{1}{2}x_2 + \frac{1}{3}\big)}{(x_2 x_1^{-1} + 1)}\big(\tfrac{1}{2}x_1 + \tfrac{1}{3}\big) \le c_*. \tag{SM5.3}$$

Take $f \in H_X$. First, for $S = \varnothing$ and $S = \{1,2\}$ we trivially have

$$\int (v^{ME}(S; X, f))^2 P_{X_S}(dx_S) \le \int f^2(x) P_X(dx).$$

Next, let $S = \{1\}$. Then, using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\int (v^{ME}(\{1\}; X, f))^2 P_{X_1}(dx_1) = \int_0^1 \bigg(\int_0^1 f(x_1, x_2) p_2(x_2) dx_2 \bigg)^2 p_1(x_1) dx_1 \\
&\le \int_0^1 \bigg(\int_0^1 1 dx_2 \bigg) \bigg(\int_0^1 f^2(x_1, x_2) p_2^2(x_2) dx_2 \bigg) p_1(x_1) dx_1 \\
&= \int_0^1 \int_0^1 f^2(x_1, x_2) \frac{p_1(x_1) p_2^2(x_2)}{p(x_1, x_2)} p(x_1, x_2) dx_1 dx_2 \\
&\le c_* \int_{[0,1]^2} f^2(x) p(x) dx = c_* \int f^2(x) P_X(dx).
\end{aligned}$$

Following the above steps, and using (SM5.3), we obtain for $S = \{2\}$

$$\begin{aligned}
&\int (v^{ME}(\{2\}; X, f))^2 P_{X_2}(dx_2) \le \int_0^1 \int_0^1 f^2(x_1, x_2) \frac{p_1^2(x_1) p_2(x_2)}{p(x_1, x_2)} p(x_1, x_2) dx_1 dx_2 \\
&\le c_* \int f^2(x) P_X(dx).
\end{aligned}$$

Thus, we established that the marginal game is bounded for every $S \subseteq N$ on H_X. Hence, the corresponding game value map $f \mapsto h[N, v^{ME}(\cdot; X, f)]$ is a bounded operator on H_X.

Finally, for $S \in \{\{1\}, \{2\}\}$ the Radon-Nikodym derivatives satisfy Lebesgue-a.e.

$$r_{\{1\}}(x_1, x_2) = r_{\{2\}}(x_1, x_2) = \frac{d[P_{X_1} \otimes P_{X_2}]}{dP_X} = \mathbb{1}_{(0,1)^2} \cdot \frac{p_1(x_1)p_2(x_2)}{p(x_1, x_2)} \to \infty$$

as $x_1, x_2 \to 0^+$, where we used (SM5.1). Thus, $r_{\{1\}}$ and $r_{\{2\}}$ are Lebesgue-essentially unbounded on $(0,1)^2$. Since the density $p(x) > 0$ on $(0,1)^2$, we conclude that $r_{\{1\}}$ and $r_{\{2\}}$ are P_X-essentially unbounded. This also means that $r = \frac{d\tilde{P}_X}{dP_X}$ is P_X-essentially unbounded.

SM6. On game value extensions to non-cooperative games. In this subsection, we discuss possible extensions of generic linear game values, which are not necessarily in the form (3.1), to non-cooperative games such as marginal and conditional games associated with an ML model. Recall that a cooperative game with n players is a set function v that acts on a finite set of players $N \subset \mathbb{N}$ and satisfies $v(\varnothing) = 0$. Typically, $N = \{1, 2, \dots, n\}$; see Appendix A.

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

$$V = \{(N, v) : \ v(\varnothing) \in \mathbb{R}, \quad v(S) = \tilde{v}(S), \ S \subseteq N, \quad |S| \geq 1, \ \text{for some } (N, \tilde{v}) \in V_0\}. \tag{SM6.1}$$

One way to construct an extension of a linear game value to V is to incorporate the value $v(\varnothing)$ into the extension itself. In what follows, for each $v \in V$, the cooperative game $\tilde{v}$ denotes its projection onto V_0 as in (SM6.1) (it agrees with v on non-empty sets). Given a linear game value h, we seek an extension $\bar{h}$ to V that satisfies:

(E1) $\bar{h}[N, \tilde{v}] = h[N, \tilde{v}]$ for $(N, \tilde{v}) \in V_0$,

(E2) $\bar{h}$ is linear on V.

Lemma SM6.1. *Let h be a linear game value. An extension $\bar{h}$ satisfying (E1)-(E2) has the representation:*

$$\bar{h}_i[N, v] = h_i[N, \tilde{v}] + \gamma_i v(\varnothing), \quad i \in N = \{1, 2, \dots, n\}, \tag{SM6.2}$$

where $\{\gamma_i\}_{i=1}^n$ are constants that depend on N. Furthermore, any game in the form (SM6.2) satisfies properties (E1)-(E2). In addition, if h is symmetric, then $\bar{h}$ is symmetric if and only if $\gamma_i = \gamma_j$, for each $i, j \in N$.

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

Definition SM6.2. *Let h be a linear game value and $\bar{h}$ its extension. We say that $\bar{h}$ is centered if $\bar{h}[N, c] = 0$ for any constant non-cooperative game $(N, c) \in V$.*

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

Lemma SM6.3. *Let h be a linear game value and $\bar{h}$ its extension with $\gamma = \{\gamma_i\}_{i=1}^n$ as in (SM6.2). Let u denote a unit, non-cooperative game on N, that is, $u(S) = 1$ for $\forall S \subseteq N$. Then*

(i) $\bar{h}$ is centered if and only if $\gamma = -h[N, \tilde{u}]$.

(ii) $\bar{h}$ is centered if and only if $\bar{h}[N, v] = h[N, (v - v(\varnothing)u)]$.

(iii) If h has the form

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

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

$$\bar{h}_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(i, N, S)\big[v(S \cup \{i\}) - v(S)\big], \quad i \in N. \tag{SM6.3}$$

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

$$\bar{h}[N, v(\cdot\,; X, f)] = \bar{h}[N, v(\cdot\,; X, f - f_0)] + f_0 \bar{h}[N, u] \quad \textit{for} \quad v \in \{v^{CE}, v^{ME}\}. \tag{SM6.4}$$

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

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

See Appendix A for more on game values.

References.

[1] 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.

[2] KhashF. Product of marginals absolutely continuous with respect to a Borel probability measure. *MathOverflow*.

[3] Jarosław Błasiok. Functions whose product with every L^1 function is L^1 *MathOverflow*.

[4] K. R. Parthasarathy. Probability measures on metric spaces. *American Mathematical Soc.*, vol. 352, 2005.