Beyond Shapley Values

Cooperative Games for the Interpretation of Machine Learning Models

[1]**Département de mathématiques** - *Université du Québec à Montréal*
[2]**Institut Intelligence et Données** - *Université Laval*

Journées PrivSec 2025
Université du Québec à Montréal - Montréal, QC, Canada.
September 12, 2025

Marouane Il Idrissi[12], Agathe Fernandes-Machado[1], Arthur Charpentier[1]

Marouane IL IDRISSI

il_idrissi.marouane@uqam.ca - marouaneilidrissi.com

Postdoctoral Researcher (since Aug. 2024)
CANSSI Distinguished Postdoctoral Fellow (since Sept. 2025)
Associate Research Member at Obvia (since 2024)

Obvia = International Observatory on the Societal Impacts of AI and Digital Technologies

Département de Mathématiques, Université du Québec à Montréal
Institut Intelligence et Données, Université Laval

Project: *Interpretability of black-box machine learning models*
With Arthur Charpentier (UQÀM), Marie-Pier Côté (ULaval)

Research interests:

Statistical Learning • XAI • Uncertainty Quantification • Sensitivity Analysis • Probability Theory • Cooperative game theory • Functional analysis

Context

Why are ML models not widely integrated into critical systems yet?

Context

Why are ML models not widely integrated into critical systems yet?

Critical systems: *hydroelectric dams, the stock market, planes, cars, the human body...*

Context

Why are ML models not widely integrated into critical systems yet?

Critical systems: *hydroelectric dams, the stock market, planes, cars, the human body...*

Main reason: The decision-making process must be **justifiable** and **justified**.
Using, e.g., statistical arguments

Context

Why are ML models not widely integrated into critical systems yet?

Critical systems: *hydroelectric dams, the stock market, planes, cars, the human body...*

Main reason: The decision-making process must be **justifiable** and **justified**.
Using, e.g., statistical arguments

However, **eXplainable AI (XAI) methods** are often backed by **empirical arguments**
"SOTA" methods, testing on limited benchmarks...

Context

Why are ML models not widely integrated into critical systems yet?

Critical systems: *hydroelectric dams, the stock market, planes, cars, the human body...*

Main reason: The decision-making process must be **justifiable** and **justified**.
Using, e.g., statistical arguments

However, **eXplainable AI (XAI) methods** are often backed by **empirical arguments**
"SOTA" methods, testing on limited benchmarks...

This is **not enough** to convince safety/regulatory authorities...

Our position:
There needs to be a theoretical justifications to the choice of an XAI method

Context

In this talk:

Revisit **post-hoc model-agnostic** XAI methods based on **cooperative game theory**

Context

In this talk:
Revisit **post-hoc model-agnostic** XAI methods based on **cooperative game theory**

They promise a variable **influence quantification** of **non-linear** ML models
☞ **Not always "well-defined", especially with dependent covariates**

Context

In this talk:
Revisit **post-hoc model-agnostic** XAI methods based on **cooperative game theory**

They promise a variable **influence quantification** of **non-linear** ML models
☞ **Not always "well-defined", especially with dependent covariates**

They heavily rely on the notion of **the Shapley values**
☞ **Often misunderstood and misused**

Context

In this talk:
Revisit **post-hoc model-agnostic** XAI methods based on **cooperative game theory**

They promise a variable **influence quantification** of **non-linear** ML models
☞ **Not always "well-defined", especially with dependent covariates**

They heavily rely on the notion of **the Shapley values**
☞ **Often misunderstood and misused**

How to use cooperative game theory right to extract insights on the behavior of black-box models?

Context

In this talk:
Revisit **post-hoc model-agnostic** XAI methods based on **cooperative game theory**

They promise a variable **influence quantification** of **non-linear** ML models
☞ **Not always "well-defined", especially with dependent covariates**

They heavily rely on the notion of **the Shapley values**
☞ **Often misunderstood and misused**

How to use cooperative game theory right to extract insights on the behavior of black-box models?

Side quests: Understand **what the Shapley values are** and how to go **beyond them**

Cooperative Game Theory

"Cooperative game theory = The art of sharing a cake"

Cooperative Game Theory

"Cooperative game theory = The art of sharing a cake"

Two ingredients:

- $D = \{1, \ldots, d\}$, a **set of players**
  The power-set $\mathcal{P}_D$ is the **set of coalitions of players**
- $v : \mathcal{P}_D \to \mathbb{R}$, a **value function**
  It **assigns a value to each coalition**

☞ (D, v) formally defines a **cooperative game**
☞ $v(D)$ is the value of the "grand coalition" (the cake)

"Cooperative game theory = The art of sharing a cake"

Two ingredients:

- $D = \{1, \ldots, d\}$, a **set of players**
  The power-set $\mathcal{P}_D$ is the **set of coalitions of players**
- $v : \mathcal{P}_D \to \mathbb{R}$, a **value function**
  It **assigns a value to each coalition**

☞ (D, v) formally defines a **cooperative game**
☞ $v(D)$ is the value of the "grand coalition" (the cake)

Main concern of the theory of cooperative games:

How can to redistribute $v(D)$ to each of the d players?

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

Data: covariates $X = (X_1, \ldots, X_d)$, and target Y
Model: estimated linear regression model $\widehat{f}(X) = \widehat{\beta}_0 + X^\top \widehat{\beta}$

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

Data: covariates $X = (X_1, \ldots, X_d)$, and target Y
Model: estimated linear regression model $\widehat{f}(X) = \widehat{\beta}_0 + X^\top \widehat{\beta}$

Cooperative game:
Players: the covariates $(X_1, \ldots, X_d)$ (rather, their indices D)
Coalitions: for $A \in \mathcal{P}_D$, the subset of covariates X_A
Value function: $v(A) = R^2_Y(X_A)$ (the R^2 coefficient of the linear model only using X_A)

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

Data: covariates $X = (X_1, \ldots, X_d)$, and target Y
Model: estimated linear regression model $\widehat{f}(X) = \widehat{\beta}_0 + X^\top \widehat{\beta}$

Cooperative game:
Players: the covariates $(X_1, \ldots, X_d)$ (rather, their indices D)
Coalitions: for $A \in \mathcal{P}_D$, the subset of covariates X_A
Value function: $v(A) = R^2_Y(X_A)$ (the R^2 coefficient of the linear model only using X_A)

☞ **The cake:** $v(D)$, the R^2 of the **full model**

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

Data: covariates $X = (X_1, \ldots, X_d)$, and target Y
Model: estimated linear regression model $\widehat{f}(X) = \widehat{\beta}_0 + X^\top \widehat{\beta}$

Cooperative game:
Players: the covariates $(X_1, \ldots, X_d)$ (rather, their indices D)
Coalitions: for $A \in \mathcal{P}_D$, the subset of covariates X_A
Value function: $v(A) = R^2_Y(X_A)$ (the R^2 coefficient of the linear model only using X_A)

☞ **The cake:** $v(D)$, the R^2 of the **full model**

Evaluate the **value function** $\forall A \in \mathcal{P}_D \iff$ The R^2 of all the 2^d nested linear models
For 20 covariate: more than a million R^2 coefficients

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

Data: covariates $X = (X_1, \ldots, X_d)$, and target Y
Model: estimated linear regression model $\widehat{f}(X) = \widehat{\beta}_0 + X^\top \widehat{\beta}$

Cooperative game:
Players: the covariates $(X_1, \ldots, X_d)$ (rather, their indices D)
Coalitions: for $A \in \mathcal{P}_D$, the subset of covariates X_A
Value function: $v(A) = R^2_Y(X_A)$ (the R^2 coefficient of the linear model only using X_A)

☞ **The cake:** $v(D)$, the R^2 of the **full model**

Evaluate the **value function** $\forall A \in \mathcal{P}_D \iff$ The R^2 of all the 2^d nested linear models
For 20 covariate: more than a million R^2 coefficients

☞ **This is way too much information to process...**

A famous example - LMG indices

Example: Lindeman, Merenda, and Gold (1980) indices

Data: covariates $X = (X_1, \ldots, X_d)$, and target Y
Model: estimated linear regression model $\widehat{f}(X) = \widehat{\beta}_0 + X^\top \widehat{\beta}$

Cooperative game:
Players: the covariates $(X_1, \ldots, X_d)$ (rather, their indices D)
Coalitions: for $A \in \mathcal{P}_D$, the subset of covariates X_A
Value function: $v(A) = R^2_Y(X_A)$ (the R^2 coefficient of the linear model only using X_A)

☞ **The cake:** $v(D)$, the R^2 of the **full model**

Evaluate the **value function** $\forall A \in \mathcal{P}_D \iff$ The R^2 of all the 2^d nested linear models
For 20 covariate: more than a million R^2 coefficients

☞ **This is way too much information to process...**

Is it possible to aggregate this information (2^d coefficients) into something more manageable?

Allocations as an aggregation

This is **exactly the role of an allocation**!
It summarizes the 2^d evaluations of v into **one quantity for each player**

Allocations as an aggregation

This is **exactly the role of an allocation**!
It summarizes the 2^d evaluations of v into **one quantity for each player**

It is a mapping $\phi : D \to \mathbb{R}$, that must ideally respect **one** criteria:

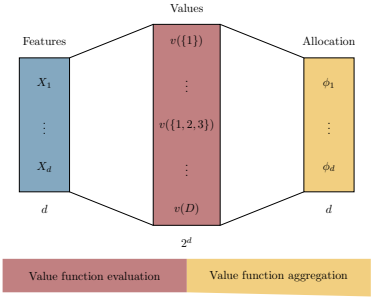
- **Efficiency:** $\sum_{i \in D} \phi(i) = v(D)$
  ☞ Ensures that **the we actually redistribute the cake**

Allocations as an aggregation

This is **exactly the role of an allocation**!
It summarizes the 2^d evaluations of v into **one quantity for each player**

It is a mapping $\phi : D \to \mathbb{R}$, that must ideally respect **one** criteria:

- **Efficiency:** $\sum_{i \in D} \phi(i) = v(D)$
  - ☞ Ensures that **the we actually redistribute the cake**



In a nuthsell:

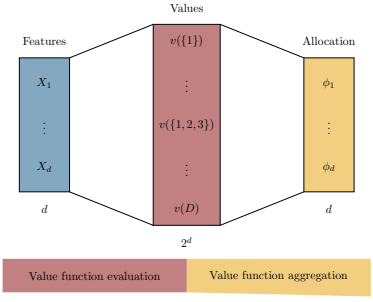
- Start with a learned model with d input features
- Chose a **value function** resulting in 2^d quantities
- Aggregate the 2^d quantities into d quantities using an **efficient allocation**

Allocations as an aggregation

This is **exactly the role of an allocation**!
It summarizes the 2^d evaluations of v into **one quantity for each player**

It is a mapping $\phi : D \to \mathbb{R}$, that must ideally respect **one** criteria:

- **Efficiency:** $\sum_{i \in D} \phi(i) = v(D)$
  - ☞ Ensures that **the we actually redistribute the cake**



In a nuthsell:

- Start with a learned model with d input features
- Chose a **value function** resulting in 2^d quantities
- Aggregate the 2^d quantities into d quantities using an **efficient allocation**

Is there a way to define efficient allocations?

Allocations as a dividend sharing mechanism

The **Harsanyi (1963) dividends** of a cooperative game (D, v) are defined as:

$$\mathcal{D}_v(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B), \quad \text{or equivalently,} \quad \mathcal{D}_v(A) = v(A) - \sum_{B \in \mathcal{P}_A} \mathcal{D}_v(B)$$

Think of them as **the added-value of a coalition**: $\mathcal{D}_v(12) = v(12) - v(1) - v(2)$

Allocations as a dividend sharing mechanism

The **Harsanyi (1963) dividends** of a cooperative game (D, v) are defined as:

$$\mathcal{D}_v(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B), \quad \text{or equivalently,} \quad \mathcal{D}_v(A) = v(A) - \sum_{B \in \mathcal{P}_A} \mathcal{D}_v(B)$$

Think of them as **the added-value of a coalition**: $\mathcal{D}_v(12) = v(12) - v(1) - v(2)$

The **Harsanyi set** is a family of **efficient allocations** that **aggregate of the Harsanyi dividends**:

$$\phi(i) = \sum_{A \in \mathcal{P}_D \,:\, i \in A} \lambda_i(A) \mathcal{D}_v(A), \quad \text{where} \quad \begin{cases} \forall i \in D, \forall A \in \mathcal{P}_D, \ \lambda_i(A) \geq 0, \\ \forall A \in \mathcal{P}_D, \ \sum_{i \in D} \lambda_i(A) = 1 \end{cases}$$

parametrized by the **weight system** $\lambda : D \times \mathcal{P}_D \to \mathbb{R}$

Allocations as a dividend sharing mechanism

The **Harsanyi (1963) dividends** of a cooperative game (D, v) are defined as:

$$\mathcal{D}_v(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B), \quad \text{or equivalently,} \quad \mathcal{D}_v(A) = v(A) - \sum_{B \in \mathcal{P}_A} \mathcal{D}_v(B)$$

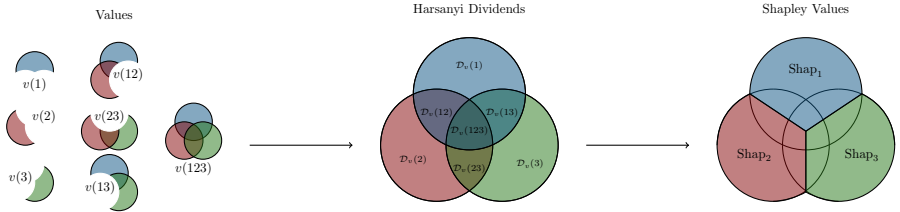
Think of them as **the added-value of a coalition**: $\mathcal{D}_v(12) = v(12) - v(1) - v(2)$

The **Harsanyi set** is a family of **efficient allocations** that **aggregate of the Harsanyi dividends**:

$$\phi(i) = \sum_{A \in \mathcal{P}_D \,:\, i \in A} \lambda_i(A) \mathcal{D}_v(A), \quad \text{where} \quad \begin{cases} \forall i \in D, \forall A \in \mathcal{P}_D,\ \lambda_i(A) \geq 0, \\ \forall A \in \mathcal{P}_D,\ \sum_{i \in D} \lambda_i(A) = 1 \end{cases}$$

parametrized by the **weight system** $\lambda : D \times \mathcal{P}_D \to \mathbb{R}$

In this setting, the **Shapley values are the egalitarian redistribution**, i.e., $\lambda_i(A) = 1/|A|$



Allocations using random orders

The **Weber (1988) set of allocations** relies on the notion of **random orders**

Allocations using random orders

The **Weber (1988) set of allocations** relies on the notion of **random orders**

Let $\mathcal{S}_D$ be the **set of permutations** $\pi = (\pi_1, \ldots, \pi_d)$ (i.e., orders) of players
For any $i \in D$, denote $\pi(i)$ the **position of player** i **in the permutation** π (i.e., $\pi_{\pi(i)} = i$)

Allocations using random orders

The **Weber (1988) set of allocations** relies on the notion of **random orders**

Let $\mathcal{S}_D$ be the **set of permutations** $\pi = (\pi_1, \ldots, \pi_d)$ (i.e., orders) of players
For any $i \in D$, denote $\pi(i)$ the **position of player** i **in the permutation** π (i.e., $\pi_{\pi(i)} = i$)

The **Weber (1988) set** is a family of **efficient allocations** as an average over the orderings

$$\begin{aligned}\phi(i) &= \mathbb{E}_{\pi \sim p}\left[v\left(\{\pi_1, \ldots, \pi_{\pi(i)}\}\right) - v\left(\{\pi_1, \ldots, \pi_{\pi(i)-1}\}\right)\right] \\ &= \sum_{\pi \in \mathcal{S}_D} p(\pi)\left[v\left(\{\pi_1, \ldots, \pi_{\pi(i)}\}\right) - v\left(\{\pi_1, \ldots, \pi_{\pi(i)-1}\}\right)\right]\end{aligned}$$

parametrized by a **probability mass function** p over the permutations $\mathcal{S}_D$.

Allocations using random orders

The **Weber (1988) set of allocations** relies on the notion of **random orders**

Let $\mathcal{S}_D$ be the **set of permutations** $\pi = (\pi_1, \ldots, \pi_d)$ (i.e., orders) of players
For any $i \in D$, denote $\pi(i)$ the **position of player** i **in the permutation** π (i.e., $\pi_{\pi(i)} = i$)

The **Weber (1988) set** is a family of **efficient allocations** as an average over the orderings

$$\begin{aligned}\phi(i) &= \mathbb{E}_{\pi \sim p}\left[v\left(\{\pi_1, \ldots, \pi_{\pi(i)}\}\right) - v\left(\{\pi_1, \ldots, \pi_{\pi(i)-1}\}\right)\right] \\ &= \sum_{\pi \in \mathcal{S}_D} p(\pi)\left[v\left(\{\pi_1, \ldots, \pi_{\pi(i)}\}\right) - v\left(\{\pi_1, \ldots, \pi_{\pi(i)-1}\}\right)\right]\end{aligned}$$

parametrized by a **probability mass function** p over the permutations $\mathcal{S}_D$.

In this setting, the **Shapley values are the uniform distribution over the permutations**, i.e., $p(\pi) = 1/d!$

$$\mathrm{Shap}(i) = \frac{1}{d!} \sum_{\pi \in \mathcal{S}_D}\left[v\left(\{\pi_1, \ldots, \pi_{\pi(i)}\}\right) - v\left(\{\pi_1, \ldots, \pi_{\pi(i)-1}\}\right)\right]$$

The recipe

Overall blueprint for using cooperative games for XAI:
(I., Charpentier, and Fernandes Machado 2025)

1. **Step 1: Identify a quantity of interest**
   Choose **a cake worth cutting**, e.g., point predictions $f(x)$, model variance $\mathbb{V}(f(X))$...
   ☞ Guides the **interpretation of the extracted insights**

The recipe

Overall blueprint for using cooperative games for XAI:
(I., Charpentier, and Fernandes Machado 2025)

1. **Step 1: Identify a quantity of interest**
   Choose **a cake worth cutting**, e.g., point predictions $f(x)$, model variance $\mathbb{V}(f(X))$...
   ☞ Guides the **interpretation of the extracted insights**

2. **Step 2: Pick a value function v**
   And make sure that $v(D)$ **is equal to the quantity of interest**, e.g.,
   $\mathbb{E}[f(X) \mid X_A = x_A]$ for $f(x)$, $\mathbb{V}(\mathbb{E}[f(X) \mid X_A])$ for $\mathbb{V}(f(X))$...
   ☞ **This step is the most important** (garbage in - garbage out)

The recipe

Overall blueprint for using cooperative games for XAI:
(I., Charpentier, and Fernandes Machado 2025)

1. **Step 1: Identify a quantity of interest**
   Choose **a cake worth cutting**, e.g., point predictions $f(x)$, model variance $\mathbb{V}(f(X))$...
   ☞ Guides the **interpretation of the extracted insights**

2. **Step 2: Pick a value function v**
   And make sure that $v(D)$ **is equal to the quantity of interest**, e.g.,
   $\mathbb{E}[f(X) \mid X_A = x_A]$ for $f(x)$, $\mathbb{V}(\mathbb{E}[f(X) \mid X_A])$ for $\mathbb{V}(f(X))$...
   ☞ **This step is the most important** (garbage in - garbage out)

3. **Step 3: Pick an efficient allocation**
   In order to summarize the information of the 2^d evaluations of v
   ☞ Less crucial, but can **highlight some model behavior**

Picking a value function

Ok, but **how do we pick a value function?**

Picking a value function

Ok, but **how do we pick a value function?**

A **bad choice** can lead to **misleading insights**

e.g., **correlation/concurvity identifiability issues** (Zhang, Martinelli, and John 2024), **lack of purity** Köhler, Rügamer, and Schmid (2024)...

Picking a value function

Ok, but **how do we pick a value function?**

A **bad choice** can lead to **misleading insights**

e.g., **correlation/concurvity identifiability issues** (Zhang, Martinelli, and John 2024), **lack of purity** Köhler, Rügamer, and Schmid (2024)...

$$f(X) = X_1 + X_2 + X_1X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Picking a value function

Ok, but **how do we pick a value function?**

A **bad choice** can lead to **misleading insights**

e.g., **correlation/concurvity identifiability issues** (Zhang, Martinelli, and John 2024), **lack of purity** Köhler, Rügamer, and Schmid (2024)...

$$f(X) = X_1 + X_2 + X_1 X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Quantity of interest: prediction $f(x)$; **Allocation:** Shapley values

Picking a value function

Ok, but **how do we pick a value function?**

A **bad choice** can lead to **misleading insights**

e.g., **correlation/concurvity identifiability issues** (Zhang, Martinelli, and John 2024), **lack of purity** Köhler, Rügamer, and Schmid (2024)...

$$f(X) = X_1 + X_2 + X_1X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Quantity of interest: prediction $f(x)$; **Allocation:** Shapley values

Conditional expectation

$$v(A) = \mathbb{E}\left[f(X) \mid X_A = x_A\right]$$

$$\mathcal{D}_v(1) = x_1 + \rho(x_1 + x_1^2 - 1) \quad \mathcal{D}_v(2) = x_2 + \rho(x_2 + x_2^2 - 1)$$

$$\mathcal{D}_v(12) = x_1x_2 - \rho(x_1 + x_1^2 + x_2 + x_2^2 - 1)$$

$$\text{Shap}_v(\{1\}) = x_1 + \frac{\rho}{2}(x_1 + x_1^2 - x_2 - x_2^2 - 1) + \frac{x_1x_2}{2}$$

$$\text{Shap}(\{2\}) = x_2 + \frac{\rho}{2}(x_2 + x_2^2 - x_1 - x_1^2 - 1) + \frac{x_1x_2}{2}$$

Picking a value function

Ok, but **how do we pick a value function?**

A **bad choice** can lead to **misleading insights**

e.g., **correlation/concurvity identifiability issues** (Zhang, Martinelli, and John 2024), **lack of purity** Köhler, Rügamer, and Schmid (2024)...

$$f(X) = X_1 + X_2 + X_1X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Quantity of interest: prediction $f(x)$; **Allocation:** Shapley values

Conditional expectation

$v(A) = \mathbb{E}\left[f(X) \mid X_A = x_A\right]$

$$\mathcal{D}_v(1) = x_1 + \rho(x_1 + x_1^2 - 1) \quad \mathcal{D}_v(2) = x_2 + \rho(x_2 + x_2^2 - 1)$$

$$\mathcal{D}_v(12) = x_1x_2 - \rho(x_1 + x_1^2 + x_2 + x_2^2 - 1)$$

$$\text{Shap}_v(\{1\}) = x_1 + \frac{\rho}{2}(x_1 + x_1^2 - x_2 - x_2^2 - 1) + \frac{x_1x_2}{2}$$

$$\text{Shap}(\{2\}) = x_2 + \frac{\rho}{2}(x_2 + x_2^2 - x_1 - x_1^2 - 1) + \frac{x_1x_2}{2}$$

Oblique projections

I. et al. (2025)

$$\mathcal{D}_v(1) = x_1 \quad \mathcal{D}_v(2) = x_2$$

$$\mathcal{D}_v(12) = x_1x_2$$

$$\text{Shap}_v(\{1\}) = x_1 + \frac{x_1x_2}{2}$$

$$\text{Shap}(\{2\}) = x_2 + \frac{x_1x_2}{2}$$

Picking an allocation

Ok, but **how do we pick an allocation?**

Picking an allocation

Ok, but **how do we pick an allocation?**

It is still an **open question**

☞ No metrics between allocations for interpretability purposes

Picking an allocation

Ok, but **how do we pick an allocation?**

It is still an **open question**

☞ No metrics between allocations for interpretability purposes

But, some choices can have interesting properties

Picking an allocation

Ok, but **how do we pick an allocation?**

It is still an **open question**
☞ No metrics between allocations for interpretability purposes

But, some choices can have interesting properties

Example: Proportional Marginal Effects (Herin et al. 2024)

- **Quantity of interest:** $\mathbb{V}(f(X))$
- **Value function:** $v(A) = \mathbb{E}\left[\mathbb{V}\left(f(X) \mid X_{D\setminus A}\right)\right]$
- **Allocation:** *Proportional values*

$$p(\pi) = \frac{L(\pi)}{\sum_{\sigma \in \mathcal{S}_D} L(\sigma)}, \quad L(\pi) = \exp\left(-\sum_{j \in D} \log\left(v\left(\{\pi_1, \dots \pi_{\pi(j)}\}\right)\right)\right)$$

Picking an allocation

Ok, but **how do we pick an allocation?**

It is still an **open question**
☞ No metrics between allocations for interpretability purposes

But, some choices can have interesting properties

Example: Proportional Marginal Effects (Herin et al. 2024)

- **Quantity of interest:** $\mathbb{V}(f(X))$
- **Value function:** $v(A) = \mathbb{E}\left[\mathbb{V}\left(f(X) \mid X_{D\setminus A}\right)\right]$
- **Allocation:** *Proportional values*

$$p(\pi) = \frac{L(\pi)}{\sum_{\sigma \in \mathcal{S}_D} L(\sigma)}, \quad L(\pi) = \exp\left(-\sum_{j \in D} \log\left(v\left(\{\pi_1, \dots \pi_{\pi(j)}\}\right)\right)\right)$$

Proposition *(Exogeneity detection).*

$$PME_i = 0 \iff X_i \text{ is not in the model.}$$

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**
- The **Shapley values** are **one example of allocation**, and there are **many more**

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**
- The **Shapley values** are **one example of allocation**, and there are **many more**
- The **choice of value function is crucial** for understanding the end product

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**
- The **Shapley values** are **one example of allocation**, and there are **many more**
- The **choice of value function is crucial** for understanding the end product
- **An allocation** is **an aggregation of information**

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**
- The **Shapley values** are **one example of allocation**, and there are **many more**
- The **choice of value function is crucial** for understanding the end product
- **An allocation** is **an aggregation of information**

Why you should you care: **Goal-oriented choice of allocation**

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**
- The **Shapley values** are **one example of allocation**, and there are **many more**
- The **choice of value function is crucial** for understanding the end product
- **An allocation** is **an aggregation of information**

Why you should you care: **Goal-oriented choice of allocation**

Chose $\lambda_i(A)$ or $p(\pi)$ to **avoid**:

- Fair washing
- The inference of private data from the final attributions
- Any other "good desired property"

Conclusion

Key take-aways:

- **Three ingredients** to using cooperative game for ML interpretability: The **quantity of interest**, the **value function**, and the **allocation**
- The **Shapley values** are **one example of allocation**, and there are **many more**
- The **choice of value function is crucial** for understanding the end product
- **An allocation** is **an aggregation of information**

Why you should you care: **Goal-oriented choice of allocation**

Chose $\lambda_i(A)$ or $p(\pi)$ to **avoid**:

- Fair washing
- The inference of private data from the final attributions
- Any other "good desired property"

☞ **Ditch the (arbitrary) Shapley values to propose new goal-oriented XAI method standards**

References i

Harsanyi, J. C. 1963. "A Simplified Bargaining Model for the n-Person Cooperative Game." Publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], *International Economic Review* 4 (2): 194–220. ISSN: 0020-6598. https://doi.org/10.2307/2525487. https://www.jstor.org/stable/2525487.

Herin, M., M. I., V. Chabridon, and B. Iooss. 2024. "Proportional Marginal Effects for Global Sensitivity Analysis" [in en]. *SIAM/ASA Journal on Uncertainty Quantification* 12, no. 2 (June): 667–692. ISSN: 2166-2525. https://doi.org/10.1137/22M153032X. https://epubs.siam.org/doi/10.1137/22M153032X.

I., M., Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, and Jean-Michel Loubes. 2025. "Hoeffding decomposition of functions of random dependent variables." *Journal of Multivariate Analysis* 208 (July): 105444. ISSN: 0047-259X. https://doi.org/10.1016/j.jmva.2025.105444. https://www.sciencedirect.com/science/article/pii/S0047259X25000399.

I., M., Arthur Charpentier, and Agathe Fernandes Machado. 2025. "Beyond Shapley Values: Cooperative Games for the Interpretation of Machine Learning Models." In *International Joint Conference on Artificial Intelligence (IJCAI) - Workshop on Explainable Artificial Intelligence (XAI)*. Montréal, Québec, Canada: Hendrik Baier and Tobias Huber and Mor Vered and Sarath Sreedharan and Katharina Weitz and Stylianos Loukas Vasileiou, August. https://hal.science/hal-05106257.

Köhler, D., D. Rügamer, and M. Schmid. 2024. *Achieving interpretable machine learning by functional decomposition of black-box models into explainable predictor effects*. ArXiv:2407.18650 [cs, stat], July. Accessed August 26, 2024. http://arxiv.org/abs/2407.18650.

Lindeman, R. H., P. F. Merenda, and R. Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis* [in English]. Scott, Foresman. ISBN: 978-0-673-15099-8. https://books.google.cz/books?id=-hfvAAAAMAAJ.

References ii

Weber, R. J. 1988. "Probabilistic values for games." Chap. 7 in *The Shapley value: essays in honor of Lloyd S. Shapley*, edited by A. E. Roth, 101–120. New York, NY: Cambridge University Press.

Zhang, Xinyu, Julien Martinelli, and S. T. John. 2024. "Challenges in interpretability of additive models." In *Proceedings of the XAI Workshop @ IJCAI 2024*. ArXiv:2504.10169 [cs]. arXiv. https://doi.org/10.48550/arXiv.2504.10169. http://arxiv.org/abs/2504.10169.

Thank you for your attention!

Any questions?

MAROUANEILIDRISSI.COM

We acknowledge the support of the Canadian Statistical Sciences Institute (CANSSI) and the Natural Sciences and Engineering Research Council of Canada (NSERC)

Nous reconnaissons le soutien de l'Institut Canadien des Sciences Statistiques (INCASS) et du Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG)