Winter values and Banzhaf indices for interpretability

C. Labreuche [1]

[1] Thales / cortAIx Labs
Palaiseau, France
email: christophe.labreuche@thalesgroup.com

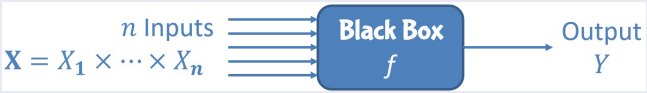
GATSBII kickoff

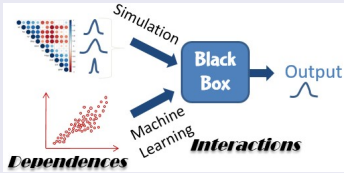
Outline

Outline

Explanation of a Black-Box

Main Approaches



Local Explanation of $f(\mathbf{x})$	• Feature Attribution (LIME, SHAP) $\phi : N \to \mathbb{R}$ with $N = \{1, \ldots, n\}$	
Global Explanation of f	• Global Sensitivity Analysis (GSA)	

Reminder on variance-based GSA (independent variables)

Hoeffding decomposition and Sobol indices

- Hoeffding decomposition

$$Y = f(\mathbf{x}) = \sum_{S \subseteq N} f_S(\mathbf{x}_S)$$

where $f_S(\mathbf{x}_S) = \sum_{K \subseteq S} (-1)^{s-k} \, \mathbb{E}_{N \setminus K} \left[f(\mathbf{X}) | \mathbf{X}_K = \mathbf{x}_K \right]$

- Sobol index

$$\mathrm{Sob}(S) = \frac{\mathrm{Var}(f_S)}{\mathrm{Var}(Y)}.$$

where

Non-negativity: $0 \leq \mathrm{Sob}(S) \leq 1 \quad \forall S \subseteq N$

Efficiency: $\sum_{S \subseteq N} \mathrm{Sob}(S) = 1$

Recent approach: Game Theory for GSA (dependent variables)

Game Theory for GSA

Two step approach:

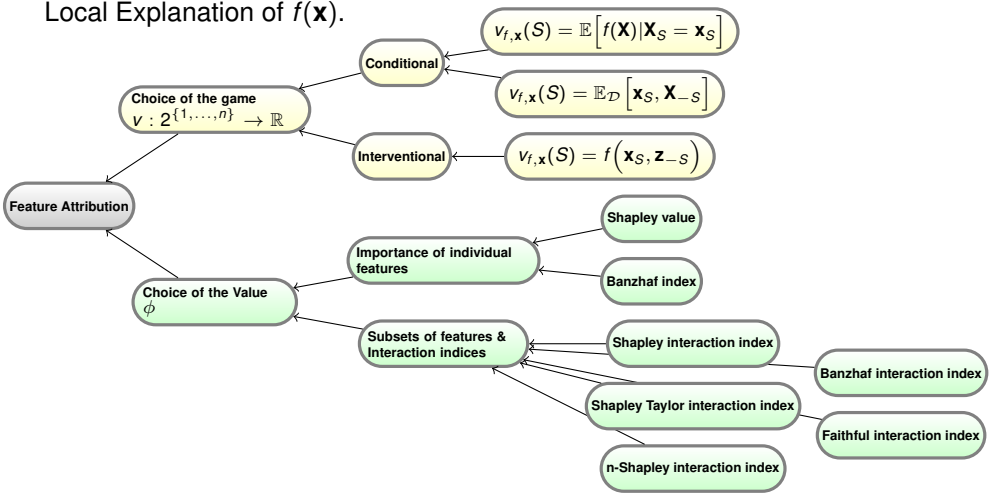
- STEP 1: GSA to construct a Cooperative Game $v : 2^N \to \mathbb{R}$.
  $v(S)$ is the contribution of subset $S \subseteq N$ to the total variance:

$$v(S) = \frac{\mathrm{Var}_S\left(\mathbb{E}_{N\setminus S}[Y|\mathbf{X}_S]\right)}{\mathrm{Var}(Y)}$$

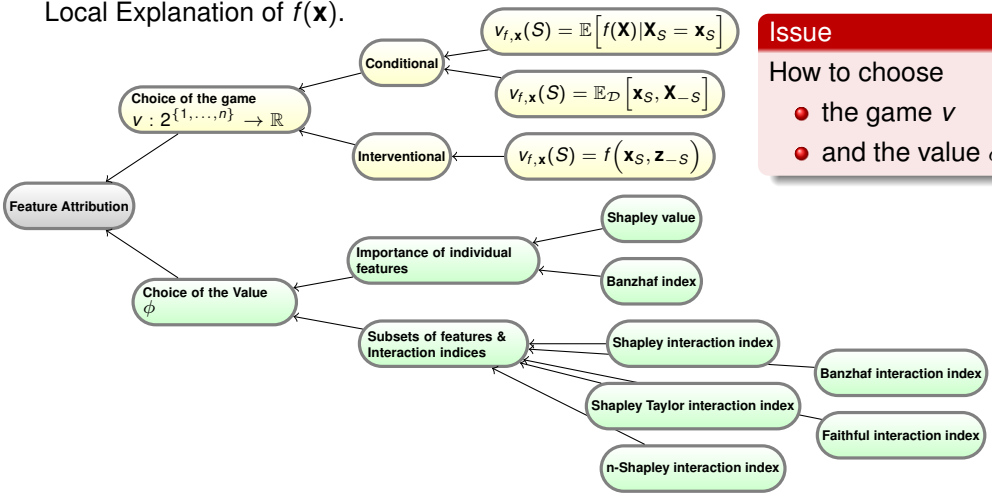
- STEP 2: Cooperative Game Theory to construct
  - importance indicators $\phi : v \mapsto \mathbb{R}^N$,
  - interaction indicators $I : v \mapsto \mathbb{R}^{2^N}$.

Overview of the Feature Attribution methods

Local Explanation of $f(\mathbf{x})$.



Overview of the Feature Attribution methods



Issue

How to choose

- the game v
- and the value ϕ?

Reminder on Shapley Effect in GSA (dependent variables)

Shapley value

$$\phi_i^{\mathrm{Sh}}(N, v) := \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \big[v(S \cup \{i\}) - v(S) \big]$$

where

Additivity: $\phi_i^{\mathrm{Sh}}(N, v+w) = \phi_i^{\mathrm{Sh}}(N, v) + \phi_i^{\mathrm{Sh}}(N, w)$

Null player: if i is null then $\phi_i^{\mathrm{Sh}}(N, v) = 0$)

Symmetry: $\forall$ permutation π on N, $\phi_{\pi(i)}^{\mathrm{Sh}}(\pi N, \pi v) = \phi_i^{\mathrm{Sh}}(N, v)$

Efficiency: $\sum_{i \in N} \phi_i^{\mathrm{Sh}}(N, v) = v(N) = 1$

Proposal

Aim

- Winter value: extension of the Shapley value to **trees**
- Proportionnal Division Effect: extion of the Shapley value to **trees** and **exclusion**
- Banzhaf indices: Hoeffding decomposition on a game

Outline

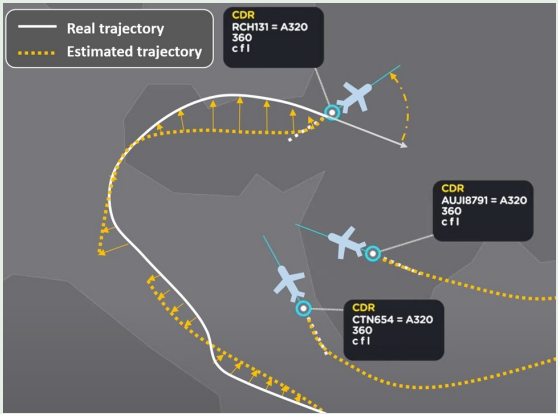
Multi-Criteria Decision problem

Multi-Criteria Decision Aiding (MCDA)

- $N = \{1, \ldots, n\}$: index set of attributes/features.
- X_i: set of values representing attribute/feature i (for $i \in N$).
- $X = X_1 \times \cdots \times X_n$: set of alternatives/instances.
  $\mathbf{x} = (x_1, \ldots, x_n) \in X$ with $x_i \in X_i$.
- Problem to solve, given a set of alternatives in X:
  - choose the *most preferred* one
  - rank the alternatives from best to worse
  - sort the alternatives into preferential categories
- $U : X \to \mathbb{R}$: utility representing preferences of decision maker over X
  - $U(\mathbf{y}) > U(\mathbf{x})$: $\mathbf{y}$ is preferred to $\mathbf{x}$
- Model U is organized in a tree

Illustration

Monitoring of Tracking System for Air Traffic Management

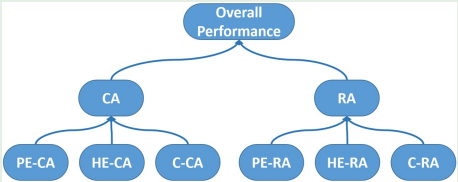


Assess the Quality of Service of a tracking system.
Tracking quality attributes:

- Position Error (PE)
- Heading Error (HE)
- Completeness (C)

Attributes are measured for each type of aircraft:

- Commercial Airplanes (CA)
- Recreational Airplanes (RA)



Outline

Outline

Idea of Influence Index

Problem at stake

In many applications, the operator seeks for simple explanations

- Indicating among the amount of input data, which one is the most influential on the decision

Aim: construct an indicator $I_i(x, y, T, U)$

- measuring the influence of factor i,
- in the comparison between two options x and y,
- in a preference model with multiple criteria represented in a tree T,
- and quantified by utility model U.

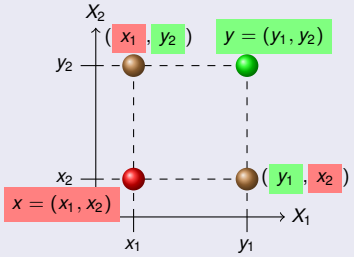
Air Traffic Management

- Why is the overall tracking QoS decreased over past 12 hours?

Axioms

Idea of the approach

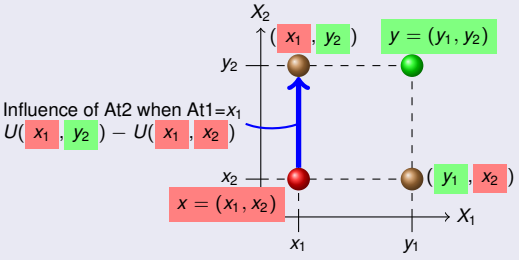
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Axioms

Idea of the approach

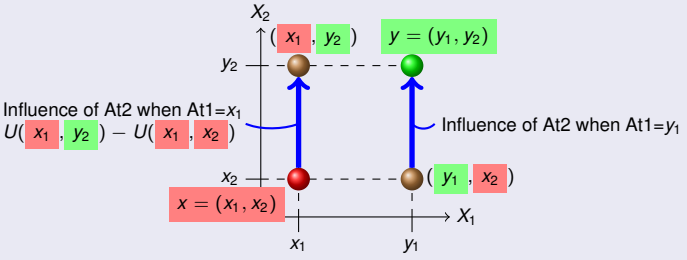
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Axioms

Idea of the approach

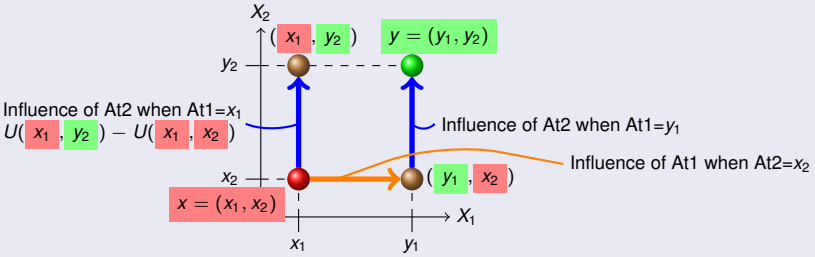
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Axioms

Idea of the approach

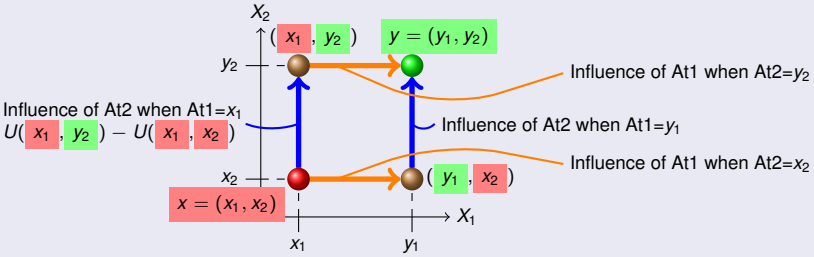
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Axioms

Idea of the approach

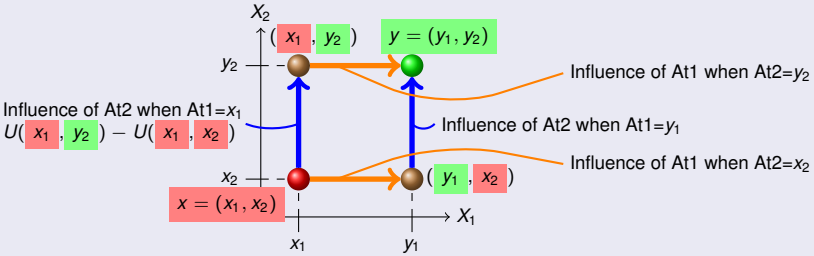
- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Axioms

Idea of the approach

- Assess the influence of a criterion in the evaluation of two alternatives x, y
- by looking at alternatives obtained by replacing subsets of values of y with values of x.
- Example with 2 attributes: $x = (x_1, x_2)$, (y_1, x_2), (x_1, y_2), and $y = (y_1, y_2)$



Restricted Value (RV)

I_k depends only on the utility U of compound options mixing values of x, y.

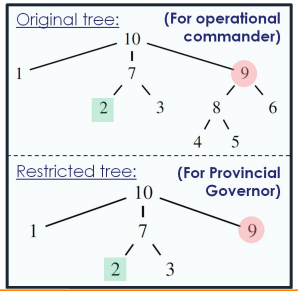
Axioms

Null Attribute (NA)

if changing x_k to y_k never changes U, then $I_k = 0$.

Consistency with Restricted Tree (CRT)

I_2 shall be the same for the original tree or a subtree where 9 becomes a leaf.



Generalized Efficiency (GE)

- General Share: $I_{10} = U(y) - U(x)$
- Decomposability: e.g. $I_9 = I_6 + I_8$

Other axioms

- Additivity (ADD): $I_k(U + U') = I_k(U) + I_k(U')$
- Restricted Equal Treatment (RET): All attributes are treated symmetrically

Are these axioms sufficient to derive I?

Theorem

There is a unique influence index satisfying **RV**, **NA**, **RET**, **ADD**, **GE** and **CRT**.

Remark

This influence index is an extension of the Shapley and Owen values on general trees.

Computational complexity

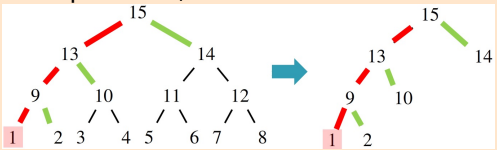
Complexity issue

Computation of I_i is exponential with n

Theorem

CRT implies that index I_i can be equivalently computed by cutting all branches not directly linking the path from node i to the root.

Example with I_1:



d	p	n	$\log_{10}\lvert\Pi(N)\rvert$	$\log_{10}\lvert\Pi(T)\rvert$	$\log_{10}\Pi(T_{\lvert J\rvert})$
2	2	4	1.38	0.903	0.602
2	3	9	5.559	3.112	1.556
2	4	16	13.320	6.901	2.76
2	5	25	25.19	12.47	4.158
2	6	36	41.57	20.0	5.715
3	2	8	4.605	2.107	0.903
3	3	27	28.036	10.115	2.334
3	4	64	89.1	28.984	4.14
3	5	125	209.27	64.454	6.237
3	6	216	412.0	122.86	8.571
4	2	16	13.3215	4.515	1.204
4	3	81	120.76	31.126	3.112
4	4	256	506.93	117.31	5.520
4	5	625	1477.7	324.35	8.316
4	6	1296	3473.0	740.04	11.429
5	2	32	35.42	9.332	1.505
5	3	243	475.76	94.156	3.89
5	4	1024	2639.7	470.65	6.901
5	5	3125	9566.3	1623.84	10.395
5	6	7776	26879	4443.15	14.286

Outline

Exclusion Principle

Motivation

In GSA, if variable *i is not in model f*, then

$$v(N \setminus \{i\}) = v(N)$$

Notation

- Dual game:
$$\overline{v}(S) = v(N) - v(N \setminus S).$$
- $v(N \setminus \{i\}) = v(N)$ iff $\overline{v}(\{i\}) = 0$.

Exclusion

If $v(\{i\}) = 0$, then $\varphi_i = 0$.

Proportional values

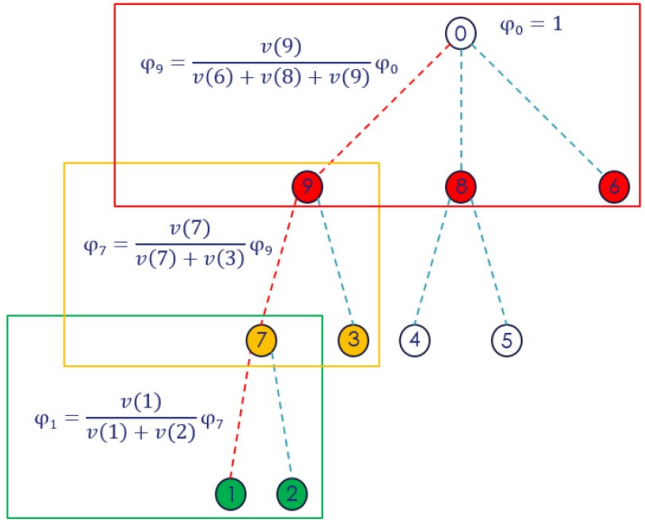
proportional division (PD)

$$\mathrm{PD}_i(N, v) = \frac{v(\{i\})}{\sum_{j \in N} v(\{j\})} v(N)$$

proportionnal Shapley value

- Möbius transform (also called Harsanyi dividend) $v(S) = \sum_{T \subseteq S} m(T)$
- Shapley value: $\mathrm{Sh}_i(N, v) = \sum_{S \subseteq N \,:\, S \ni i} \frac{1}{s}\, m(S)$
- proportionnal Shapley value: $\mathrm{PSh}_i(N, v) = \sum_{S \subseteq N \,:\, S \ni i} \frac{v(\{i\})}{\sum_{j \in S} v(\{j\})}\, m(S)$

Proportional Division Effect



Comparison

	Modèle non linéaire	Variables dépendantes	Exclusion	Efficience sur arbres
Indices de Sobol	Oui	Non	Non	Non
Shapley Effects	Oui	Oui	Non	Non
Proportionnal Marginal Effects	Oui	Oui	Oui	Non
Winter Effects	Oui	Oui	Non	Oui
Proportionnal Division Effect	Oui	Oui	Oui	Oui

Comparison

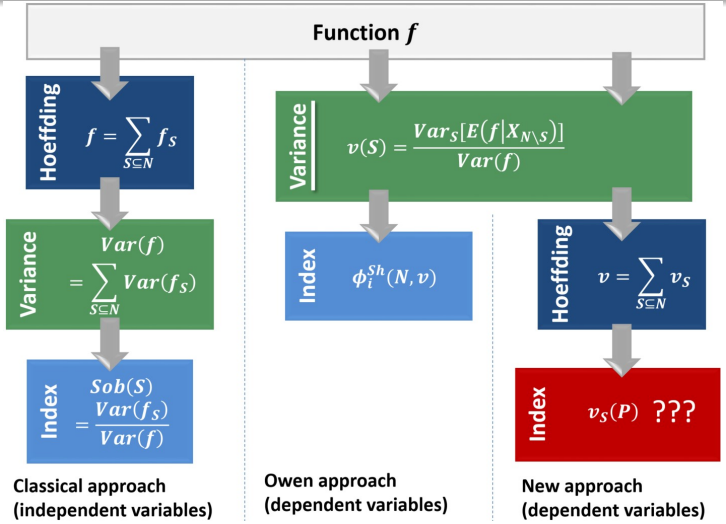
Number of used coalitions

Consider a *balanced* tree with depth q with 2 leaves.

	$q = 3$ $2^q = 8$	$q = 5$ $2^q = 32$	$q = 7$ $2^q = 128$
nombre de coalitions Shapley	256	$4,3 \times 10^9$	$3,4 \times 10^{39}$
nombre de coalitions par Winter	80	1728	30976
nombre de coalitions par PD	7	63	255

Outline

Approach


Function f
Hoeffding
$f = \sum_{S \subseteq N} f_S$
Variance
$Var(f) = \sum_{S \subseteq N} Var(f_S)$
Index
$Sob(S) = \frac{Var(f_S)}{Var(f)}$
Classical approach
(independent variables)
Variance
$v(S) = \frac{Var_S[E(f|X_{N \setminus S})]}{Var(f)}$
Index
$\phi_i^{Sh}(N, v)$
Owen approach
(dependent variables)
Hoeffding
$v = \sum_{S \subseteq N} v_S$
Index
$v_S(P)$???
New approach
(dependent variables)

Hoeffding decomposition on the game

Theorem

Assuming that the presence or absence of the variables to a coalition are i.i.d. and given by a uniform distribution, the Hoeffding decomposition of game v is given by for all $P \subseteq N$

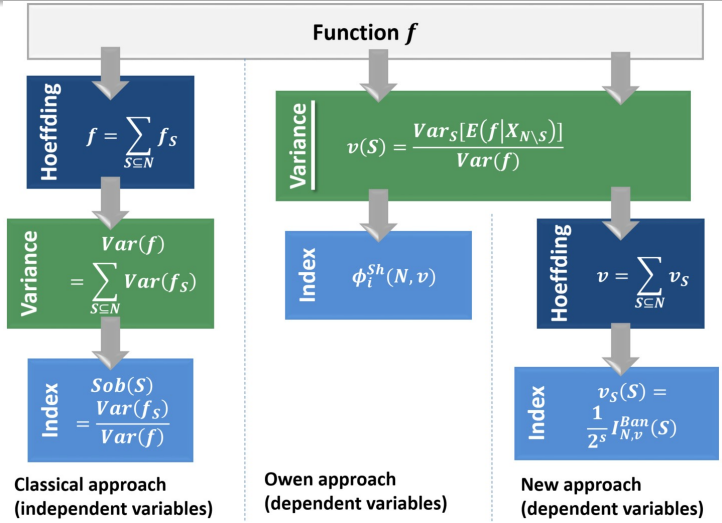
$$v(P) = \sum_{S \subseteq N} v_S(P \cap S),$$

where $v_S : 2^S \to \mathbb{R}$ is given by

$$v_S(P) = \frac{(-1)^{s-p}}{2^s} I^{\text{Ban}}_{N,v}(S)$$

$$I^{\text{Ban}}_{N,v}(S) = \frac{1}{2^{n-s}} \sum_{T \subseteq N \setminus S} \sum_{K \subseteq S} (-1)^{s-k} v(K \cup T).$$

Summary of the main result



Illustration

Example 1

- $n = 2$
- $f(x_1, x_2) = x_1$
- $\mathbf{X} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$
- $\mathrm{Var}(Y) = 1$ with $Y = f(X_1, X_2)$
- $v(\emptyset) = 0$, $v(\{1\}) = 1$, $v(\{2\}) = \rho^2$, $v(\{1,2\}) = 1$

- New Index:

$$\mathrm{Index}(\emptyset) = \frac{2+\rho^2}{4} \ , \ \mathrm{Index}(\{1,2\}) = -\frac{\rho^2}{4}$$

$$\mathrm{Index}(\{1\}) = \frac{1}{2} - \frac{\rho^2}{4} \ , \ \mathrm{Index}(\{2\}) = \frac{\rho^2}{4}$$

- Shapley effect:

$$\phi_1^{\mathrm{Sh}}(N, v) = 1 - \frac{\rho^2}{2} \ , \ \phi_2^{\mathrm{Sh}}(N, v) = \frac{\rho^2}{2}$$

Illustration

Example 2

- $n = 2$
- $f(x_1, x_2) = x_1 + \alpha\, x_1\, x_2$
- $\mathbf{X} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
- $\mathrm{Var}(Y) = 1 + \alpha^2$ with $Y = f(X_1, X_2)$
- $v(\emptyset) = 0$, $v(\{1\}) = \frac{1}{1+\alpha^2}$, $v(\{2\}) = 0$, $v(\{1,2\}) = 1$

- New Index:

$$\mathrm{Index}(\emptyset) = \frac{1 + \alpha^2/2}{2(1+\alpha^2)}\ ,\ \mathrm{Index}(\{1,2\}) = \frac{\alpha^2/2}{2(1+\alpha^2)}$$

$$\mathrm{Index}(\{1\}) = \frac{1 + \alpha^2/2}{2(1+\alpha^2)}\ ,\ \mathrm{Index}(\{2\}) = \frac{\alpha^2/2}{2(1+\alpha^2)}$$

- Shapley effect:

$$\phi_1^{\mathrm{Sh}}(N, v) = \frac{2 + \alpha^2}{2(1+\alpha^2)}\ ,\ \phi_2^{\mathrm{Sh}}(N, v) = \frac{\alpha^2}{2(1+\alpha^2)}$$

Perspectives for GATSBII