# Théorie des jeux coopératifs et Analyse de Sensibilité

## Méthodes d'estimation des effets de Shapley

[1]**EDF Lab Chatou** - Département PRISME
[2]**Institut de Mathématiques de Toulouse**
[3]**SINCLAIR AI Lab**

***Journées thématiques MEXICO***
*Jeudi 4 Mai 2022*

Marouane Il Idrissi[1,2,3], Bertrand Iooss[1,2,3], Vincent Chabridon[1,3].

# Introduction

Les *éléments conditionnels* sont définis par :

$$V_A = S_A^{\text{clos}} = \frac{\mathbb{V}\left(\mathbb{E}\left[Y \mid X_A\right]\right)}{\mathbb{V}(Y)} \quad \text{et} \quad E_A = \frac{\mathbb{E}\left[\mathbb{V}\left(Y|X_{\overline{A}}\right)\right]}{\mathbb{V}(Y)}$$

L'estimation des effets de Shapley se fait en 2 temps :

1. **Estimation des éléments conditionnels** : estimer $V_A$ ou $E_A$, pour tout $A \subseteq D$.
2. **W-Aggregation** : calculer les effets de Shapley en utilisant la formule (Song, Nelson, and Staum 2016)

$$\begin{aligned} Sh_j &= \frac{1}{d} \sum_{A \subset D \setminus \{j\}} \binom{d-1}{|A|}^{-1} \left(V_{\{A \cup \{j\}\}} - V_A\right) \\ &= \frac{1}{d} \sum_{A \subset D \setminus \{j\}} \binom{d-1}{|A|}^{-1} \left(E_{\{A \cup \{j\}\}} - E_A\right) \end{aligned}$$

# Sommaire

# Sommaire

# Estimation Monte-Carlo

Il faut être en mesure de pouvoir :

- Simuler selon la loi jointe des entrées $P_X$ ;
- Simuler selon toutes les lois marginales $P_{X_A}$, pour tout $A \subset D$ ;
- Simuler selon toutes les lois conditionnelles $P_{X|X_A}$, pour tout $A \subset D$.

On tire un échantillon $X_A$ de taille $N_o$ de la loi marginale $P_{X_A}$. Pour chaque observation $x_A^{(j)}$ de $X_A$, on tire un échantillon $\widetilde{X}_{A,i} = \left(X \mid X_A = x_A^{(i)}\right)$ de taille $N_i$ de la loi $P_{X|X_A}$.

Avec un échantillon $X$ de taille $N_v$ de la loi jointe $P_X$, on estime empiriquement la variance de $Y$ :

$$\widehat{\mathbb{V}(Y)} = \frac{1}{N_v - 1} \sum_{i=1}^{N_v} (y_i - \overline{y})^2$$

avec $y_i = G(x^{(i)})$ où $x^{(i)}$ est la $i$-ème observation de l'échantillon $X$, et $\overline{y} = \frac{1}{N_v} \sum_{i=1}^{N_v} y_i$.

# Estimation Monte Carlo

On estime ensuite $V_A = \frac{\mathbb{V}(\mathbb{E}[Y|X_A])}{\mathbb{V}(Y)}$ par leur contrepartie empirique :

$$\mathbb{V}\left(\widehat{\mathbb{E}\left[Y \mid X_A\right]}\right) = \frac{1}{N_o - 1} \sum_{i=1}^{N_o} \left( \frac{1}{N_i} \sum_{j=1}^{N_i} G(\widetilde{X}_{A,i}^{(j)}) - \overline{y} \right)^2 .$$

Ainsi, on obtient un estimateur (consistant) de $V_A$ :

$$\widehat{V_A} = \frac{\mathbb{V}\left(\widehat{\mathbb{E}\left[Y \mid X_A\right]}\right)}{\widehat{\mathbb{V}(Y)}}.$$

Il reste qu'à estimer tous les $\widehat{V_A}$ pour tout $A \subseteq D$. Cet algorithme est **très gourmand en nombre d'appel au code de calcul** : il nécéssite $(N_v + d! \times (d-1) \times N_o \times N_i)$ appels à $G$ pour calculer **tous les effets de Shapley**.

(Fonction `shapleyPermEx` du package `sensitivity`, voir le TP2)

# Estimation Monte-Carlo

Une **heuristique** pour optimiser le compromis coût/précision de l'estimateur : fixer $N_i = 3$.

Au-delà du coût d'estimation, la nécessité de **connaître** et **pouvoir simuler** selon les lois conditionnelles peut être compliqué, **surtout en situation de dépendance entre les variables**.

# Estimation Monte-Carlo

Une **heuristique** pour optimiser le compromis coût/précision de l'estimateur : fixer $N_i = 3$.

Au-delà du coût d'estimation, la nécessité de **connaître** et **pouvoir simuler** selon les lois conditionnelles peut être compliqué, **surtout en situation de dépendance entre les variables**.

Est-il possible d'estimer les effets de Shapley avec un échantillon i.i.d. des entrées/sorties du modèle ?

# Sommaire

# Estimation KNN

Démarche proposée par (Broto, Bachoc, and Depecker 2020) : dans une approche Double Monte-Carlo, **ne pas simuler selon les lois conditionnelles**, mais plutôt **approcher un échantillon des lois conditionnelles** par procédure **KNN**.

**Intuition :** Au tableau.

# Estimation KNN

Démarche proposée par (Broto, Bachoc, and Depecker 2020) : dans une approche Double Monte-Carlo, **ne pas simuler selon les lois conditionnelles**, mais plutôt **approcher un échantillon des lois conditionnelles** par procédure **KNN**.

**Intuition :** Au tableau.

Soit $\left(X^{(1)}, \dots, X^{(N)}\right)$ un $N$-échantillon i.i.d de $X$, et notons $k_N^A(l,n)$ l'indice tel que $X_A^{\left(k_N^A(l,n)\right)}$ soit le $n$-ième plus proche voisin de $X_A^{(l)}$ dans $\left(X_A^{(1)}, \dots, X_A^{(N)}\right)$.

$$\widehat{E_{A,\mathrm{KNN}}} = \frac{1}{N}\sum_{l=1}^{N}\left(\frac{1}{N_s-1}\sum_{i=1}^{N_s}\left[G\left(X^{\left(k_N^{\bar{A}}(l,i)\right)}\right) - \frac{1}{N_s}\sum_{h=1}^{N_s} G\left(X^{\left(k_N^{\bar{A}}(l,h)\right)}\right)\right]^2\right)$$

où $N_s$ est l'hyper-paramètre relatif au nombre de voisin à sélectionner. De plus, $\widehat{E}_{A,\mathrm{KNN}}$ converge vers $E_A$ en probabilité lorsque $N$ tends vers $+\infty$

(Fonctions `shapleySubsetMC`, `sobolshap_knn` et `shapleysobol_knn` du package `sensitivity`, voir le TP2)

# Estimation KNN

**Points positifs :**

- Approche *data-driven*.
- Estimateur convergent sous certaines hypothèses peu contraignantes.
- Estimation peu coûteuse.

**Points négatifs :**

- Le choix du nombre de voisins $N_s$ est un paramètre à choisir, et peut **beaucoup influencer** les résultats (manque de stabilité).

# References i


Broto, B, F Bachoc, and M Depecker. 2020. "Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution" [in en]. *SIAM/ASA Journal on Uncertainty Quantification* 8, no. 2 (January): 693–716. ISSN: 2166-2525, accessed December 2, 2020. https://doi.org/10.1137/18M1234631. https://epubs.siam.org/doi/10.1137/18M1234631.

Song, E., B. L. Nelson, and J. Staum. 2016. "Shapley Effects for Global Sensitivity Analysis: Theory and Computation" [in English]. *SIAM/ASA Journal on Uncertainty Quantification* 4, no. 1 (January): 1060–1083. ISSN: 2166-2525. https://doi.org/10.1137/15M1048070. http://epubs.siam.org/doi/10.1137/15M1048070.