A Shapley Value Proxy for Efficient Data-Set Valuation

Felipe Garrido-Lucero[1,2], Maxime Vono[3], Benjamin Heymann[3],
Patrick Loiseau[1] & Vianney Perchet[2,3]

1 INRIA FairPlay Team
2 CREST, ENSAE
3 Criteo AI Lab

From Matchings to Markets - CIRM
December 14th 2023

- Université de Lorraine, Metz, France, **24th - 26th June** 2024
- For PhD students and Postdocs
  - (Math) Roberto Cominetti: Non-Expansive Maps and Fixed Point Iterations
  - (CS) Panayotis Mertikopoulos: Online Optimization and Learning in Games
  - (Econ) Nicolas Vieille: Social Learning
- Workshop 27th and 28th June 2024
- www.gaimss24.org or info@gaimss24.org

Outline

Data-Set Valuation

Regression model and the Binning method

Cooperative game theory model

Shapley Value

Discrete Uniform Shapley

Numerical results

Conclusions

Outline

Data-Set Valuation

Data-set valuation
Quantify the contribution of players when sharing their data-sets towards solving some machine learning task

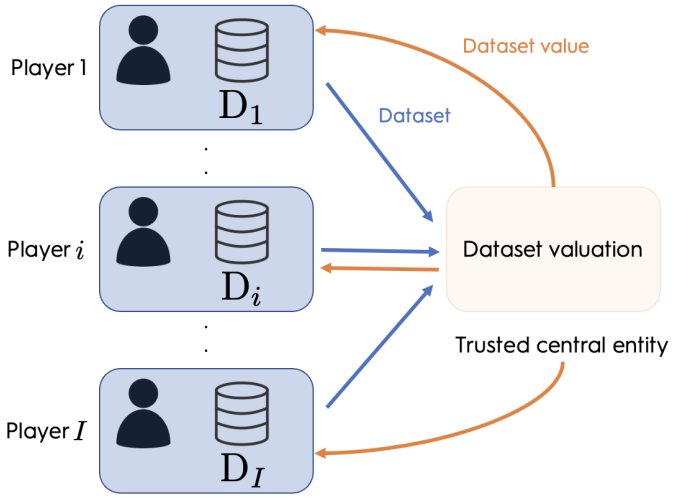


Figure: Data-set valuation problem

Applications

- Mechanism design: First step towards incentivizing data sharing
- Federated learning: Optimally/fairly agents' revenue

Approach

- Cooperative game theory
- Shapley value

Literature

- Data-set valuation problem

Agarwal et al. 2019, Sim et al. 2020, Tay et al. 2021

- Shapley value has found its place in machine learning e.g. in
  - data valuation: Jia et al. 2019, Ghorbani et al. 2020, Kwon et al. 2021, Kwon and Zou 2022, Schoch et al. 2022
  - variable selection: Cohen et al. 2005
  - feature importance: Lundberg et al. 2020, Lundberg and Lee 2017, Covert et al. 2020
  - model interpretation: Chen et al. 2019

- Shapley value approximations

Castro et al. 2009, Ghorbani and Zou 2019, Mitchell et al. 2022

Outline

Regression model

Model

- Set of agents $\mathcal{I} = \{1, ..., I\}$
- Player i has a data-set $\mathrm{D}_i \subseteq \mathcal{X} \times \mathcal{Y}$ from some distribution F_i
- Denote $n_i = |\mathrm{D}_i|$ and $N = \sum_{i \in \mathcal{I}} n_i$

Regression

$$Y_\ell = f(X_\ell) + \varepsilon_\ell,$$

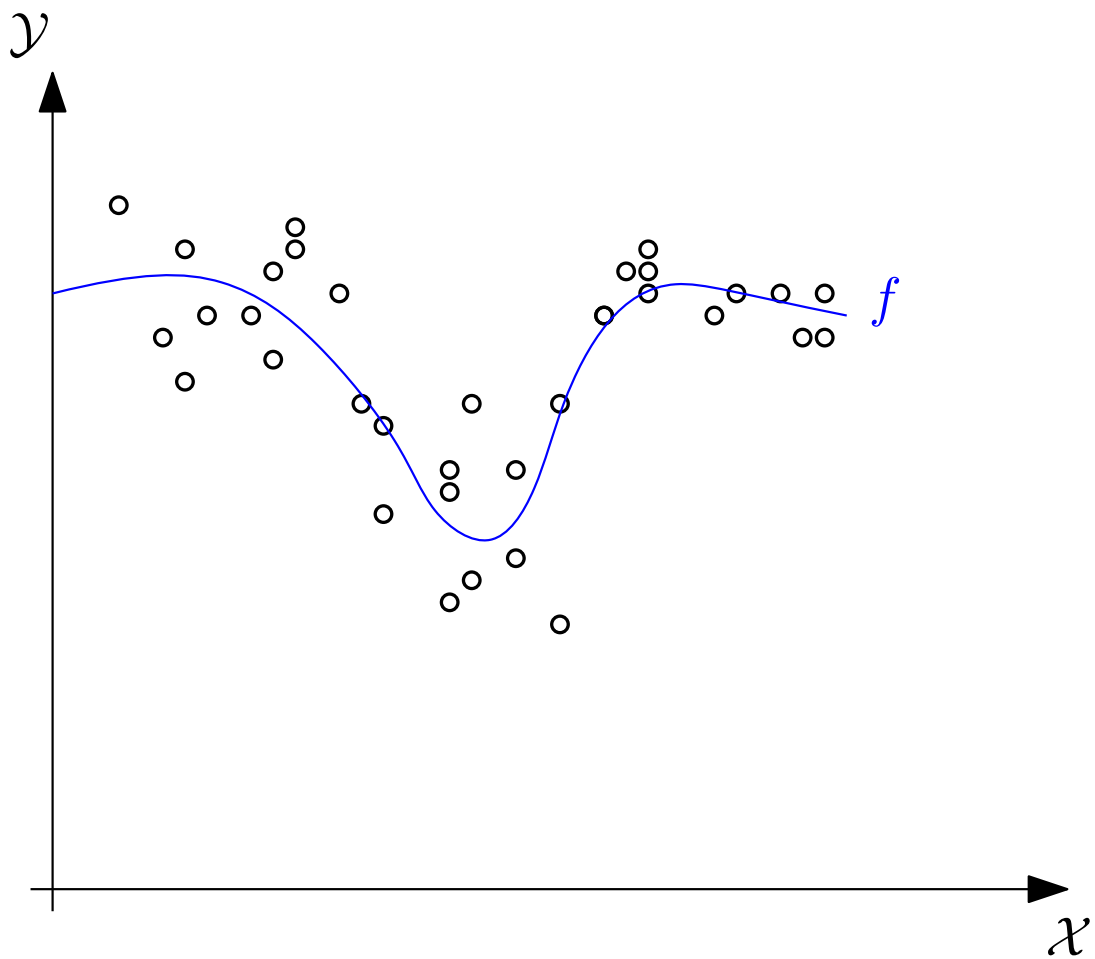
ε_ℓ is white noise and f is unknown

Goal

- To estimate $f^*(x) = \mathbb{E}[Y \mid X = x]$

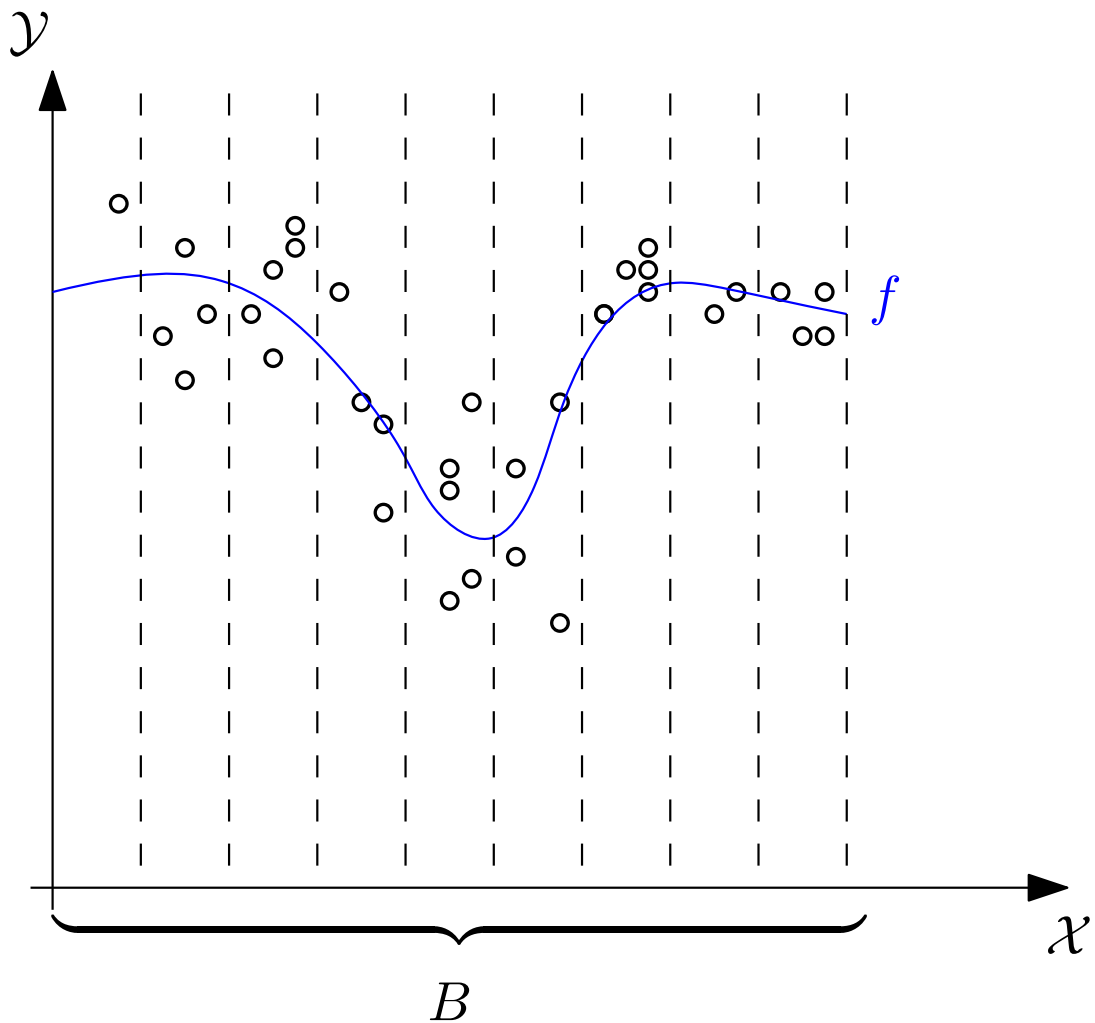
Binning method - Regressogram

Let $B \in \mathbb{N}$ be fixed



Binning method - Regressogram

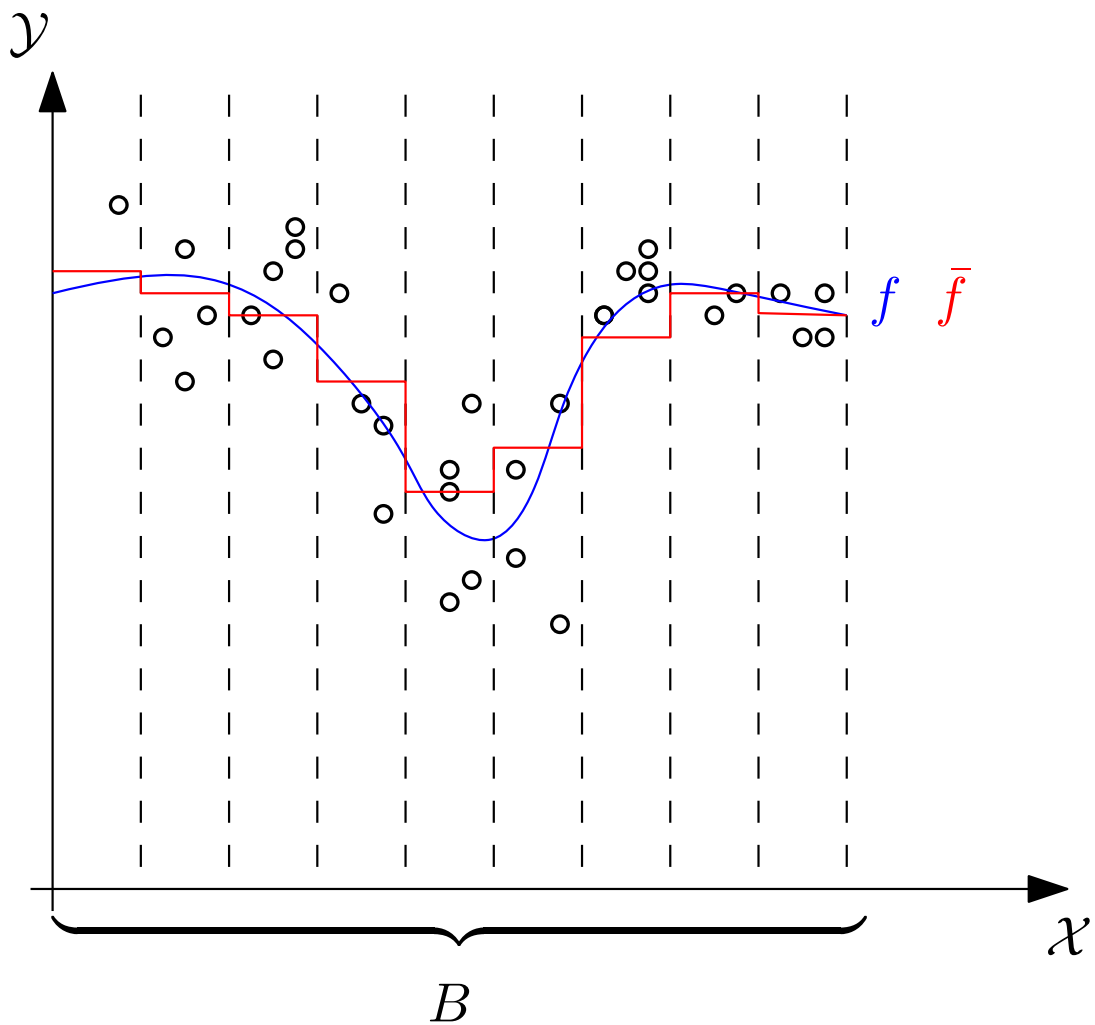
Let $B \in \mathbb{N}$ be fixed. $\mathrm{D}_i = \bigcup_{b \in [B]} \mathrm{D}_i^b$, with $\mathrm{D}_i^b \cap \mathrm{D}_i^{b'} = \varnothing$ for any two bins b and b'



Binning method - Regressogram

Risk $\longrightarrow \mathbb{E}_x\left[(\hat{f}(x) - f^*(x))^2\right] = \mathbb{E}_x\left[(\hat{f}(x) - \bar{f}(x))^2\right] + \mathbb{E}_x\left[(\bar{f}(x) - f^*(x))^2\right]$

$$\bar{f} = \sum_{b \in [B]} \bar{f}_b(x), \quad \bar{f}_b(x) = \mathbb{E}_x[f(x) \mid x \in b]$$



Estimation & Approximation error

- With this in mind,

$$\mathbb{E}\left[(\hat{f}(x) - f^*(x))^2\right] = \underbrace{\mathbb{E}\left[(\hat{f}(x) - \bar{f}(x))^2\right]}_{\text{Estimation error}} + \underbrace{\mathbb{E}\left[(\bar{f}(x) - f^*(x))^2\right]}_{\text{Approximation error}}$$

  - Estimation error $\nearrow B$ and $\searrow N$
  - Approximation error $\searrow B$

  $\Big\} \longrightarrow$ Trade-off in B

- Approximation error does not depend on $(\mathrm{D}_i)_{i\in\mathcal{I}}$ but only on B

Data-set valuation

Quantify the contribution of players when sharing their data-sets towards ~~solving some ML task~~ decreasing the estimation error

Outline

Model

- Set of players $\mathcal{I} = \{1, ..., I\}$
- Value function $v : 2^{\mathcal{I}} \to \mathbb{R}$, $S \subseteq \mathcal{I}$,

$$v(S) = -\mathbb{E}_x\left[(\hat{f}_S(x) - \bar{f}(x))^2\right] = -\sum_{b\in[B]} \mathbb{E}_x^b\left[(\hat{f}_{S,b}(x) - \bar{f}_b(x))^2\right] =: \sum_{b\in[B]} v_b(S)$$

- We measure the agents' contribution in each bin
- For a fixed bin $b \in [B]$ we suppose an homogeneous distribution,

$$v_b(S) = w_b(n_S^b), \text{ where } n_S^b = \sum_{i\in S} n_i^b$$

- In linear regression ($\mathcal{X} = \mathbb{R}^d$, $\sigma_\varepsilon = \mathbb{E}[\varepsilon_\ell^2]$),

$$w_b(n_S^b) = \frac{-d\sigma_\varepsilon^2}{n_S - (d+1)}$$

Outline

Shapley value

- Classical solution concept in cooperative game theory
- Average marginal contribution of player i to all subcoalitions $S \subseteq \mathcal{I} \setminus \{i\}$
- Given $v : 2^{\mathcal{I}} \to \mathbb{R}$, the Shapley value of player i is

$$\varphi_i(v) = \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \left[v(S \cup \{i\}) - v(S) \right]$$

Shapley value

- Classical solution concept in cooperative game theory
- Average marginal contribution of player i to all subcoalitions $S \subseteq \mathcal{I} \setminus \{i\}$
- Given $v : 2^{\mathcal{I}} \to \mathbb{R}$, the Shapley value of player i is

$$\begin{aligned}\varphi_i(v) &= \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[v(S \cup \{i\}) - v(S) \Big] \\ &= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[v_b(S \cup \{i\}) - v_b(S) \Big]\end{aligned}$$

Shapley value

- Classical solution concept in cooperative game theory
- Average marginal contribution of player i to all subcoalitions $S \subseteq \mathcal{I} \setminus \{i\}$
- Given $v : 2^{\mathcal{I}} \to \mathbb{R}$, the Shapley value of player i is

$$\begin{aligned}
\varphi_i(v) &= \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[v(S \cup \{i\}) - v(S) \Big] \\
&= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[v_b(S \cup \{i\}) - v_b(S) \Big] \\
&= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[w_b(n_S^b + n_i^b) - w_b(n_S^b) \Big]
\end{aligned}$$

Shapley value

- Classical solution concept in cooperative game theory
- Average marginal contribution of player i to all subcoalitions $S \subseteq \mathcal{I} \setminus \{i\}$
- Given $v : 2^{\mathcal{I}} \to \mathbb{R}$, the Shapley value of player i is

$$\begin{aligned}
\varphi_i(v) &= \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[v(S \cup \{i\}) - v(S) \Big] \\
&= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[v_b(S \cup \{i\}) - v_b(S) \Big] \\
&= \sum_{b \in [B]} \frac{1}{I} \sum_{S \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|S|}} \Big[w_b(n_S^b + n_i^b) - w_b(n_S^b) \Big] \\
&=: \sum_{b \in [B]} \varphi_i^b(w_b)
\end{aligned}$$

- We compute each local Shapley value $\varphi_i^b(w_b)$ (intractable)

Outline

Local Shapley value

- Take $B = 1$. The (local) Shapley value can be rewritten as,

$$\varphi_i(w) = \mathbb{E}_{K \sim \mathrm{U}(\{0,1,\dots,I-1\})} \left[\mathbb{E}_{S \sim \mathrm{U}\left(2_K^{\mathcal{I}\setminus\{i\}}\right)} \left[w(n_S + n_i) - w(n_S) \right] \right]$$

- What is the distribution of $(n_S)_{S \subseteq \mathcal{I}\setminus\{i\}}$?

- Take $B = 1$. The (local) Shapley value can be rewritten as,

$$\varphi_i(w) = \mathbb{E}_{K\sim \mathrm{U}(\{0,1,\dots,I-1\})}\left[\mathbb{E}_{S\sim \mathrm{U}\left(2_K^{\mathcal{I}\setminus\{i\}}\right)}\left[w(n_S + n_i) - w(n_S)\right]\right]$$

- What is the distribution of $(n_S)_{S\subseteq\mathcal{I}\setminus\{i\}}$?

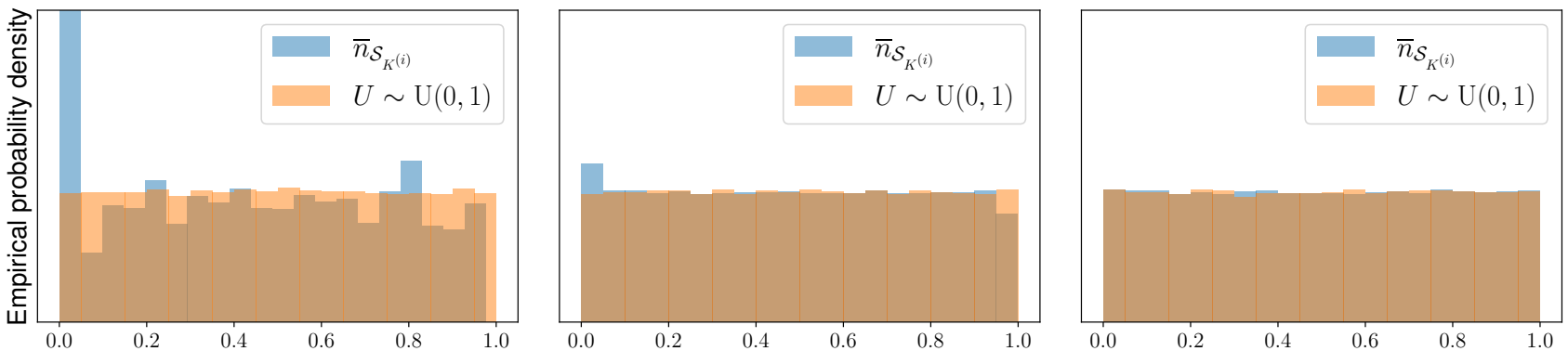


Figure: (left) $I = 10$, (middle) $I = 50$, (right) $I = 500$. 10^5 samples for each random variable and a number of data points per player drawn from $\mathrm{U}([100])$. $\bar{n}_{\mathcal{S}_K}$ stands for $n_{\mathcal{S}_K}$ normalised.

Discrete uniform Shapley

Theorem Let $n_{S_K} := \sum_{j \in S_K} n_j$, where $S_K \sim \mathrm{U}\left(2_K^{\mathcal{I}\setminus\{i\}}\right)$ and $K \sim \mathrm{U}(\{0, ..., I-1\})$. Then,

$$\frac{n_{S_K}}{\sum_{j \in \mathcal{I}\setminus\{i\}} n_j} \xrightarrow{I\to\infty} \mathrm{U}([0,1])$$

Discrete uniform Shapley

Theorem Let $n_{S_K} := \sum_{j \in S_K} n_j$, where $S_K \sim \mathrm{U}\left(2_K^{\mathcal{I}\setminus\{i\}}\right)$ and $K \sim \mathrm{U}(\{0, ..., I-1\})$. Then,

$$\frac{n_{S_K}}{\sum_{j \in \mathcal{I}\setminus\{i\}} n_j} \xrightarrow{I\to\infty} \mathrm{U}([0,1])$$

Definition Discrete uniform Shapley

$$\psi_i := \frac{1}{I}\sum_{k=0}^{I-1}[w(k\mu_{-i} + n_i) - w(k\mu_{-i})], \quad \mu_{-i} := \frac{1}{I-1}\sum_{j \in \mathcal{I}\setminus\{i\}} n_j$$

Discrete uniform Shapley

Theorem Let $n_{S_K} := \sum_{j \in S_K} n_j$, where $S_K \sim \mathrm{U}\left(2_K^{\mathcal{I}\setminus\{i\}}\right)$ and $K \sim \mathrm{U}(\{0, ..., I-1\})$. Then,

$$\frac{n_{S_K}}{\sum_{j \in \mathcal{I}\setminus\{i\}} n_j} \xrightarrow{I\to\infty} \mathrm{U}([0,1])$$

Definition Discrete uniform Shapley

$$\psi_i := \frac{1}{I}\sum_{k=0}^{I-1}[w(k\mu_{-i} + n_i) - w(k\mu_{-i})], \quad \mu_{-i} := \frac{1}{I-1}\sum_{j \in \mathcal{I}\setminus\{i\}} n_j$$

Theorem Whenever $w \in \mathcal{C}^2$ is increasing and $|w''(x)|x^2 \le w_\infty$,

$$|\varphi_i - \psi_i| \le \frac{w_\infty}{2I\mu_{-i}}\left(9(1+\ln(I))\sigma_{-i}^2 + 2R_{-i}^2\tau_{-i}\right)$$

$$R_{-i} = \max_{j \in \mathcal{I}\setminus\{i\}} |n_j - \mu_{-i}|, \quad \tau_{-i} = \max_{j \in \mathcal{I}\setminus\{i\}} n_j / \min_{j \in \mathcal{I}\setminus\{i\}} n_j$$

- Linear regression satisfies assumptions

Outline

DU-Shapley vs Monte Carlo based methods

- DU-Shapley ψ_i needs I function valuations
- Monte Carlo $\hat{\varphi}_i$, to achieve $\mathbb{P}(|\varphi_i(w) - \hat{\varphi}_i(w)| \leq \varepsilon) \geq 1 - \delta$, needs

$$T_{\text{perm}}(\varepsilon, \delta) = \frac{2r_v^2 I}{\varepsilon^2} \log\left(\frac{2I}{\delta}\right), \quad r_v := \max_{S_1, S_2 \subseteq \mathcal{I}} \{v(S_1) - v(S_2)\}$$

DU-Shapley vs Monte Carlo based methods

- DU-Shapley ψ_i needs I function valuations

- Monte Carlo $\hat{\varphi}_i$, to achieve $\mathbb{P}(|\varphi_i(w) - \hat{\varphi}_i(w)| \leq \varepsilon) \geq 1 - \delta$, needs

$$T_{\text{perm}}(\varepsilon, \delta) = \frac{2r_v^2 I}{\varepsilon^2} \log\left(\frac{2I}{\delta}\right), \quad r_v := \max_{S_1, S_2 \subseteq \mathcal{I}} \{v(S_1) - v(S_2)\}$$

- Fixing $T_{\text{perm}} = I$,

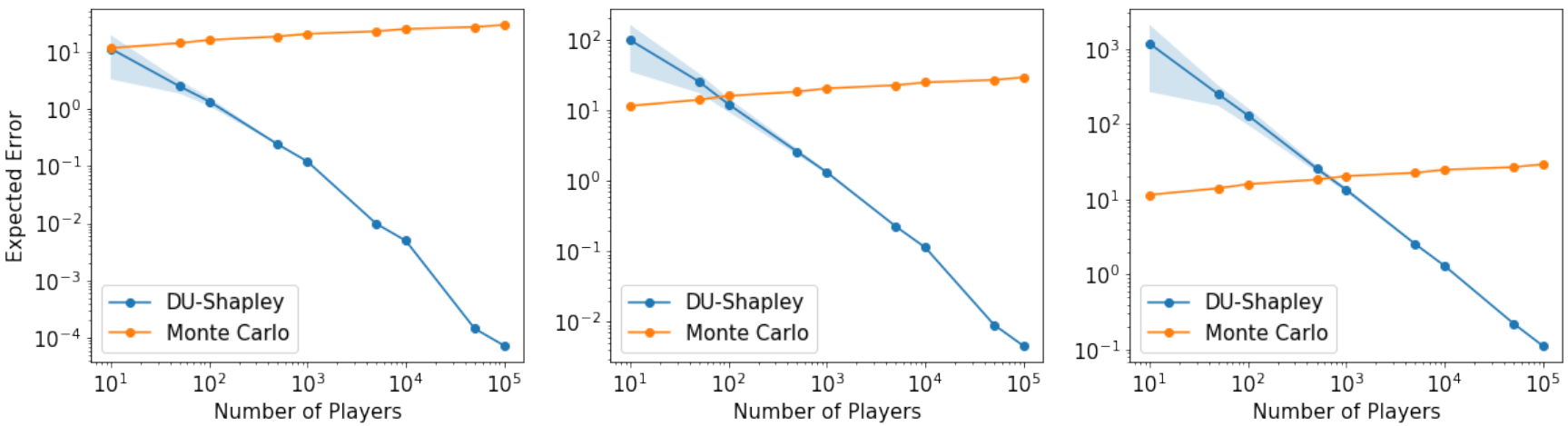
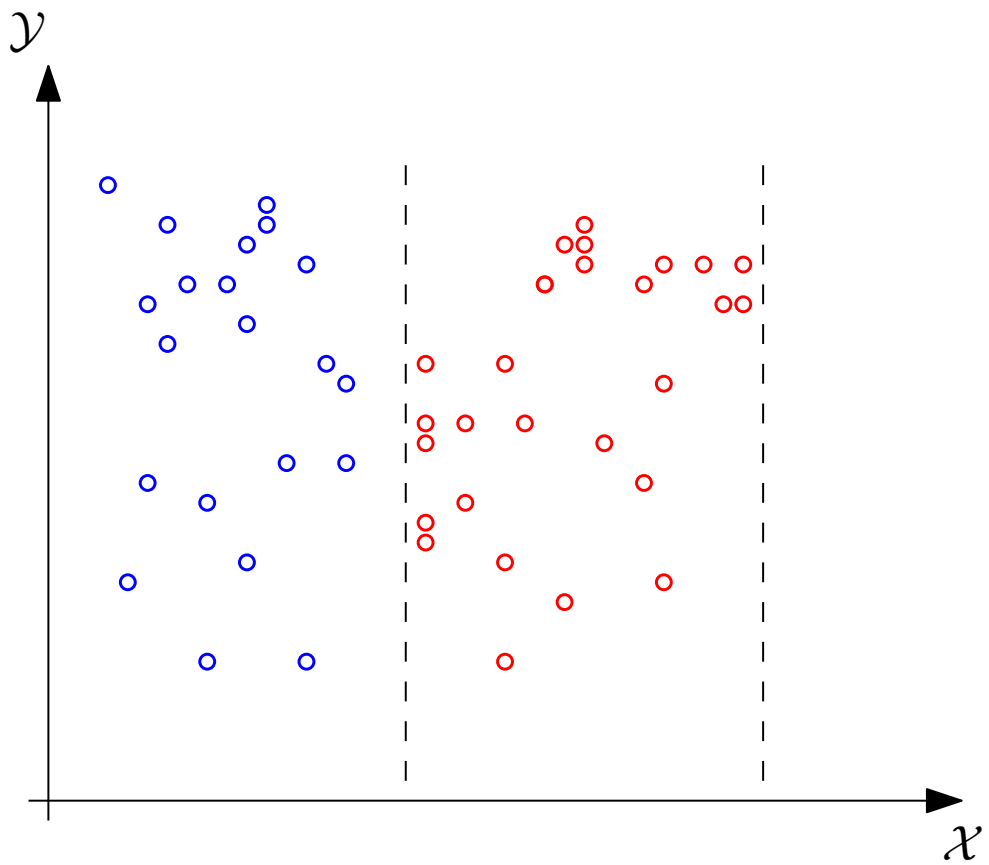


Figure: For each value of I, we drew 100 times the data points of each player from $\mathrm{U}([n_{\max}])$, with (left) $n_{\max} = 10^2$, (center) $n_{\max} = 10^3$, and (right) $n_{\max} = 10^4$.

DU-Shapley vs Monte Carlo based methods (2)



Linear regression $\longrightarrow$ $w_b(n_S^b) = \dfrac{d\sigma_\varepsilon^2}{d+1-n_S^b}$ $\quad w_r(n_S^r) = \dfrac{d\sigma_\varepsilon^2}{d+1-n_S^r}$

DU-Shapley vs Monte Carlo based methods (3)

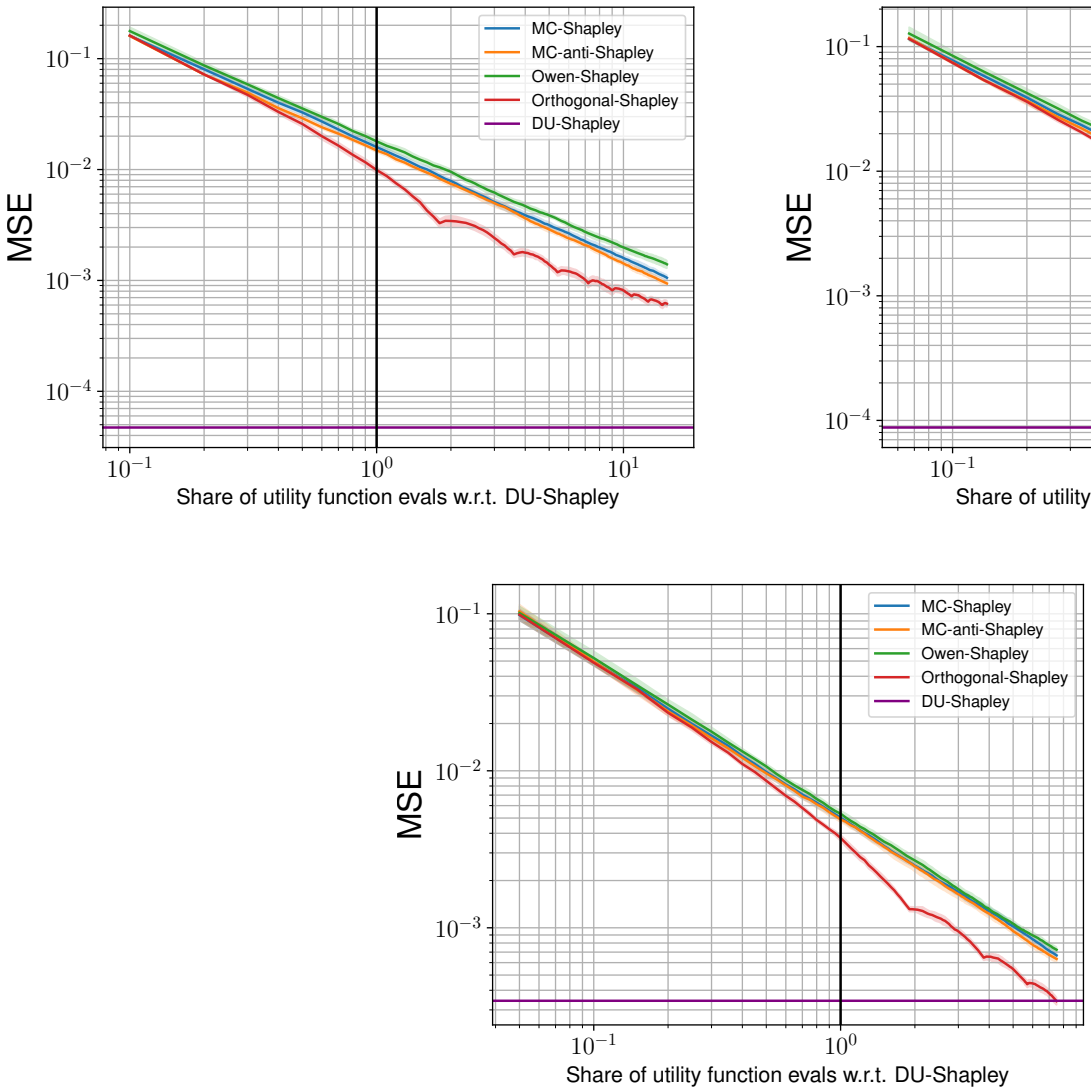


Figure: DU-Shapley vs MC-based approximations on synthetic datasets. $I = 10$, $I = 15$, and $I = 20$. $n_i^b, n_i^r \sim \mathrm{U}(\{20, \ldots, 10^3\})$.

Outline

Conclusions

Conclusions

- We study data-set valuation problem
- Mixing ML and GT we model this problem as the sum of cooperative games
- We have an efficient Shapley value approximation
- We have theoretical guarantees for our method
- Our method outperforms state of the art Monte Carlo approximation schemes

Future work

- Study the heterogeneity per bin
- Design mechanism to incentivise data-sharing

Thanks
www.gaimss24.org