DU-Shapley: A Shapley Value Proxy for Efficient Data-Set Valuation

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

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

Workshop on Blockchain and Decentralized Finance
September 21th 2023

Outline

Outline

Data-set valuation

Quantify incremental contribution of players by sharing their data-sets with other players towards solving some ML task

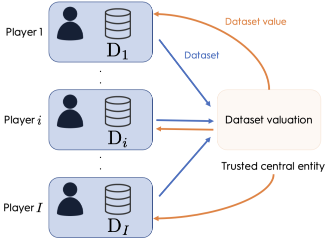

Figure: Data-set valuation problem

> **Data-set valuation**
> Quantify incremental contribution of players by sharing their data-sets with other players towards solving some ML task

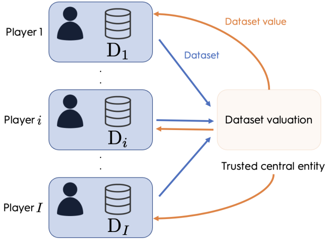

Figure: Data-set valuation problem

- First step towards incentivise parties to share data
- Cooperative game theory fits this framework
- One of the most studied solution concept is the Shapley value

Outline

Model - Linear regression

- Set of players $\mathcal{I} = \{1, ..., I\}$
- Player i has a data-set

$$\mathrm{D}_i = \{(x_i^{(j)}, y_i^{(j)})\}_{j \in [n_i]}$$

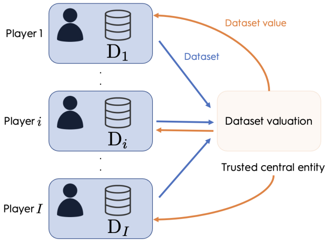

Figure: Data-set valuation problem

- Set of players $\mathcal{I} = \{1, ..., I\}$
- Player i has a data-set

$$\mathrm{D}_i = \{(x_i^{(j)}, y_i^{(j)})\}_{j \in [n_i]}$$

- Linear regression

$$Y_i = X_i \theta + \eta_i, \eta_i \sim \mathrm{N}(0_{n_i}, \varepsilon_i^2 \mathrm{I}_{n_i})$$

$$x_i^{(j)} \sim p_X^{(i)}, \text{ for any } j \in [n_i], \text{ and } \varepsilon_i \sim p_\varepsilon$$

where $\theta \in \mathbb{R}^d$ is a ground-truth parameter

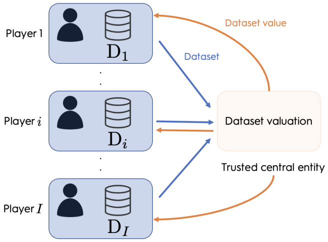

Figure: Data-set valuation problem

Model - Linear regression

- Set of players $\mathcal{I} = \{1, ..., I\}$
- Player i has a data-set

$$\mathrm{D}_i = \{(x_i^{(j)}, y_i^{(j)})\}_{j \in [n_i]}$$

- Linear regression

$$Y_i = X_i\theta + \eta_i, \eta_i \sim \mathrm{N}(0_{n_i}, \varepsilon_i^2 \mathrm{I}_{n_i})$$

$x_i^{(j)} \sim p_X^{(i)}$, for any $j \in [n_i]$, and $\varepsilon_i \sim p_\varepsilon$

where $\theta \in \mathbb{R}^d$ is a ground-truth parameter

- Value function $u : 2^{\mathcal{I}} \to \mathbb{R}$, $\forall \mathcal{S} \subseteq \mathcal{I}$,

$$u(\mathcal{S}) = v(\{\mathrm{D}_i\}_{i \in \mathcal{S}})$$

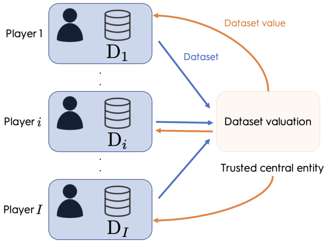

Figure: Data-set valuation problem

Model - Linear regression

- Set of players $\mathcal{I} = \{1, ..., I\}$
- Player i has a data-set

$$\mathrm{D}_i = \{(x_i^{(j)}, y_i^{(j)})\}_{j \in [n_i]}$$

- Linear regression

$$Y_i = X_i \theta + \eta_i, \eta_i \sim \mathrm{N}(0_{n_i}, \varepsilon_i^2 \mathrm{I}_{n_i})$$

$$x_i^{(j)} \sim p_X^{(i)}, \text{ for any } j \in [n_i], \text{ and } \varepsilon_i \sim p_\varepsilon$$

where $\theta \in \mathbb{R}^d$ is a ground-truth parameter

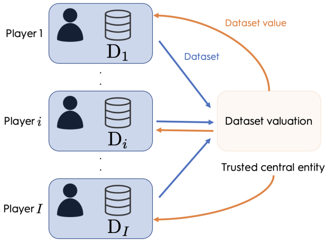

Figure: Data-set valuation problem

- Value function $u : 2^{\mathcal{I}} \to \mathbb{R}$, $\forall \mathcal{S} \subseteq \mathcal{I}$,

$$u(\mathcal{S}) = v(\{\mathrm{D}_i\}_{i \in \mathcal{S}}) = -\mathbb{E}\left[\left(x^\top \theta - x^\top \hat{\theta}_{\mathcal{S}}\right)^2\right]$$

where $\hat{\theta}_{\mathcal{S}} = (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^\top Y_{\mathcal{S}}$ is the maximum likelihood estimator and $x \sim p_X^{\text{test}}$

Outline

Shapley value

- Classical solution concept in cooperative game theory
- Average marginal contribution of player i to all subcoalitions $\mathcal{S} \subseteq \mathcal{I} \setminus \{i\}$

Shapley value

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

$$\varphi_i(u) = \qquad \Big[u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})\Big]$$

Shapley value

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

$$\varphi_i(u) = \qquad \frac{1}{\binom{I-1}{|\mathcal{S}|}} \Big[u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) \Big]$$

Shapley value

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

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

Shapley value

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

$$\begin{aligned}\varphi_i(u) &= \frac{1}{I} \sum_{\mathcal{S} \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|\mathcal{S}|}} \Big[u(\mathcal{S} \cup \{i\}) - u(\mathcal{S})\Big] \\ &= \frac{1}{I!} \sum_{\pi \in \Pi(\mathcal{I})} \Big[u(\mathcal{P}_i^{\pi} \cup \{i\}) - u(\mathcal{P}_i^{\pi})\Big]\end{aligned}$$

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

$$\begin{aligned}\varphi_i(u) &= \frac{1}{I} \sum_{\mathcal{S} \subseteq \mathcal{I} \setminus \{i\}} \frac{1}{\binom{I-1}{|\mathcal{S}|}} \left[u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) \right] \\ &= \frac{1}{I!} \sum_{\pi \in \Pi(\mathcal{I})} \left[u(\mathcal{P}_i^{\pi} \cup \{i\}) - u(\mathcal{P}_i^{\pi}) \right]\end{aligned}$$

- The intractability of the Shapley value obliges to study approximation schemes
- Castro, Gómez, and Tejada (2009) proposed a Monte Carlo approximation

$$\hat{\varphi}_i(u) = \frac{1}{T} \sum_{t=1}^{T} \left[u(\mathcal{P}_i^{\pi_t} \cup \{i\}) - u(\mathcal{P}_i^{\pi_t}) \right]$$

Outline

Homogeneous case

- Assumption: Take $p_X^{(i)} = p_X, \forall i \in \mathcal{I}$

- Assumption: Take $p_X^{(i)} = p_X, \forall i \in \mathcal{I}$
- Whenever $x \sim p_X$, it holds,

$$u(\mathcal{S}) = -\mathbb{E}\left[\left(x^\top \theta - x^\top \hat{\theta}_{\mathcal{S}}\right)^2\right] = \frac{d\sigma_\varepsilon^2}{d+1-n_{\mathcal{S}}} = w(n_{\mathcal{S}})$$

where $n_{\mathcal{S}} := \sum_{i \in \mathcal{S}} n_i$

- Assumption: Take $p_X^{(i)} = p_X, \forall i \in \mathcal{I}$
- Whenever $x \sim p_X$, it holds,

$$u(\mathcal{S}) = -\mathbb{E}\left[\left(x^\top \theta - x^\top \hat{\theta}_{\mathcal{S}}\right)^2\right] = \frac{d\sigma_\varepsilon^2}{d+1-n_{\mathcal{S}}} = w(n_{\mathcal{S}})$$

where $n_{\mathcal{S}} := \sum_{i \in \mathcal{S}} n_i$

- Having this in mind, the Shapley value can be rewritten as,

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

- Assumption: Take $p_X^{(i)} = p_X, \forall i \in \mathcal{I}$
- Whenever $x \sim p_X$, it holds,

$$u(\mathcal{S}) = -\mathbb{E}\left[\left(x^\top \theta - x^\top \hat{\theta}_{\mathcal{S}}\right)^2\right] = \frac{d\sigma_\varepsilon^2}{d+1-n_{\mathcal{S}}} = w(n_{\mathcal{S}})$$

where $n_{\mathcal{S}} := \sum_{i \in \mathcal{S}} n_i$

- Having this in mind, the Shapley value can be rewritten as,

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

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

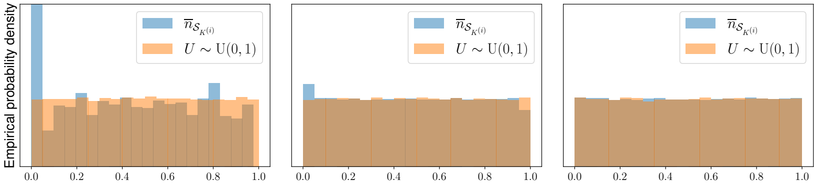

Figure: (left) $I = 10$, (middle) $I = 50$, (right) $I = 500$. 10^5 samples for each random variable and a number of data points per player drawn from $\mathrm{U}([100])$. $\bar{n}_{\mathcal{S}_K}$ stands for $n_{\mathcal{S}_K}$ normalised.

Homogeneous case (2)

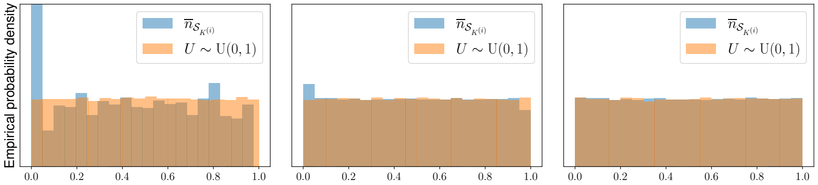

Figure: (left) $I = 10$, (middle) $I = 50$, (right) $I = 500$. 10^5 samples for each random variable and a number of data points per player drawn from $\mathrm{U}([100])$. $\bar{n}_{\mathcal{S}_K}$ stands for $n_{\mathcal{S}_K}$ normalised.

Theorem

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

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

Outline

- Discrete uniform Shapley

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

$$\mu_{-i} := \frac{1}{I-1} \sum_{j \in \mathcal{I} \setminus \{i\}} n_j$$

- Discrete uniform Shapley

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

Theorem

Under mild conditions,

$$|\varphi_i - \psi_i| \le f\left(\mu_{-i}, \sigma_{-i}, w(n_{\mathcal{I}\setminus\{i\}}), R_{-i}, n_{-i}^{\text{max}}\right) \cdot \frac{\ln(I-1)}{I-1}$$

where $\sigma_{-i}^2 = \frac{1}{I-1}\sum_{j\in\mathcal{I}\setminus\{i\}}(n_j-\mu_{-i})^2$, $R_{-i} := \max_{j\in\mathcal{I}\setminus\{i\}}|n_j-\mu_{-i}|$, and $n_{-i}^{\max} := \max_{j\in\mathcal{I}\setminus\{i\}} n_j$.

Outline

DU-Shapley vs Monte Carlo based methods

- The number of permutations T s.t. $\mathbb{P}(|\varphi_i(w) - \hat{\varphi}_i(w)| \leq \varepsilon) \geq 1 - \delta$ is,

$$T_{\mathsf{perm}}(\varepsilon, \delta) = \frac{2r_u^2 I}{\varepsilon^2} \log\left(\frac{2I}{\delta}\right), \quad r_u := \max_{S_1, S_2 \subseteq \mathcal{I}} \{u(S_1) - u(S_2)\}.$$

- The number of permutations T s.t. $\mathbb{P}(|\varphi_i(w) - \hat{\varphi}_i(w)| \leq \varepsilon) \geq 1 - \delta$ is,

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

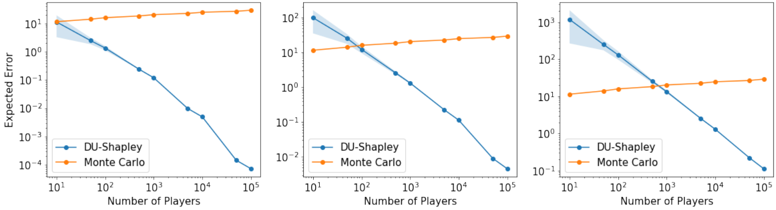

Figure: Monte Carlo's expected error for limited sampling budget $(T = I)$ versus `DU-Shapley`'s expected bias. For each value of I, we drew 100 times the data points of each player from $\mathrm{U}([n_{\max}])$, with (left) $n_{\max} = 10^2$, (center) $n_{\max} = 10^3$, and (right) $n_{\max} = 10^4$.

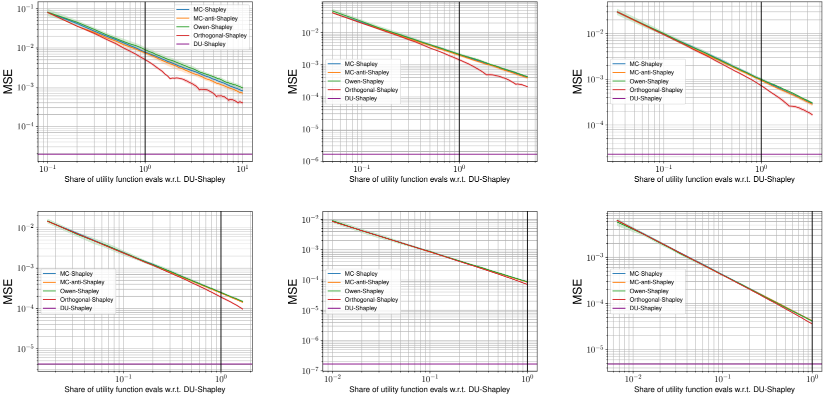

Figure: DU-Shapley vs MC-based approximations on synthetic datasets. Constant number of utility function evaluations equal to I, illustrated by the vertical black line, From left to right, (top) $I = 10$, $I = 20$ and $I = 30$, (bottom) $I = 60$, $I = 100$ and $I = 150$. Dataset size drawn from the Uniform distribution $\mathrm{U}(\{20, \dots, 10^3\})$.

Outline

Heterogeneous case

- Assumption: Take $p_X^{(i)} = \mathrm{N}(0, \sigma_i \eta^2 \mathrm{I}_d)$

Heterogeneous case

- Assumption: Take $p_X^{(i)} = \mathrm{N}(0, \sigma_i \eta^2 \mathrm{I}_d)$
- There is no close formula for $u(\mathcal{S})$ anymore. However, for $x \sim \mathrm{N}(0, \eta \mathrm{I}_d)$,

$$u(\mathcal{S}) \approx \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_\mathcal{S}, \sigma_\mathcal{S})}, \quad q(n_\mathcal{S}, \sigma_\mathcal{S}) = \left\lfloor \frac{\left(\sum_{i \in \mathcal{S}} \sigma_i n_i\right)^2}{\sum_{i \in \mathcal{S}} \sigma_i^2 n_i} \right\rfloor$$

Heterogeneous case

- Assumption: Take $p_X^{(i)} = \mathrm{N}(0, \sigma_i \eta^2 \mathrm{I}_d)$
- There is no close formula for $u(\mathcal{S})$ anymore. However, for $x \sim \mathrm{N}(0, \eta \mathrm{I}_d)$,

$$u(\mathcal{S}) \approx \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_{\mathcal{S}}, \sigma_{\mathcal{S}})}, \quad q(n_{\mathcal{S}}, \sigma_{\mathcal{S}}) = \left\lfloor \frac{\left(\sum_{i \in \mathcal{S}} \sigma_i n_i\right)^2}{\sum_{i \in \mathcal{S}} \sigma_i^2 n_i} \right\rfloor$$

- The Shapley value becomes,

$$\varphi_i(u) = \mathbb{E}_{K \sim \mathrm{U}([I-1\}])} \left[\mathbb{E}_{\mathcal{S} \sim \mathrm{U}\left(\left[2_K^{\mathcal{I} \setminus \{i\}}\right]\right)} \left[u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) \right] \right]$$

$$u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) = \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_{\mathcal{S} \cup \{i\}}, \sigma_{\mathcal{S} \cup \{i\}})} - \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_{\mathcal{S}}, \sigma_{\mathcal{S}})}$$

- Assumption: Take $p_X^{(i)} = \mathrm{N}(0, \sigma_i \eta^2 \mathrm{I}_d)$
- There is no close formula for $u(\mathcal{S})$ anymore. However, for $x \sim \mathrm{N}(0, \eta \mathrm{I}_d)$,

$$u(\mathcal{S}) \approx \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_\mathcal{S}, \sigma_\mathcal{S})}, \quad q(n_\mathcal{S}, \sigma_\mathcal{S}) = \left\lfloor \frac{\left(\sum_{i \in \mathcal{S}} \sigma_i n_i\right)^2}{\sum_{i \in \mathcal{S}} \sigma_i^2 n_i} \right\rfloor$$

- The Shapley value becomes,

$$\varphi_i(u) = \mathbb{E}_{K \sim \mathrm{U}([I-1\}])} \left[\mathbb{E}_{\mathcal{S} \sim \mathrm{U}\left(\left[2_K^{\mathcal{I} \setminus \{i\}}\right]\right)} \left[u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) \right] \right]$$

$$u(\mathcal{S} \cup \{i\}) - u(\mathcal{S}) = \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_{\mathcal{S} \cup \{i\}}, \sigma_{\mathcal{S} \cup \{i\}})} - \frac{d\sigma_\varepsilon^2}{d + 1 - q(n_\mathcal{S}, \sigma_\mathcal{S})}$$

- What is the distribution of $(q(n_{\mathcal{S} \cup \{i\}}, \sigma_{\mathcal{S} \cup \{i\}}))_{\mathcal{S} \subseteq \mathcal{I} \setminus \{i\}}$ and $(q(n_\mathcal{S}, \sigma_\mathcal{S}))_{\mathcal{S} \subseteq \mathcal{I} \setminus \{i\}}$?

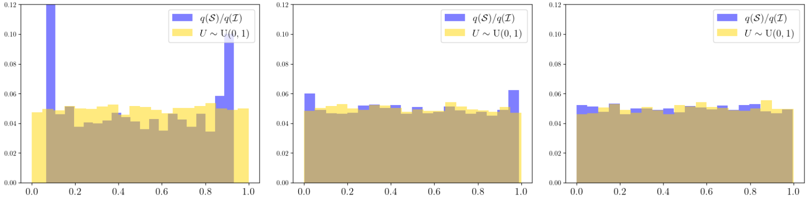

Figure: (left) $I = 10$, (middle) $I = 50$, (right) $I = 500$. We considered 10^4 samples for each random variable, a number of data points per player drawn from $\mathrm{U}([100])$, and $\sigma_i \sim \mathrm{U}([10])$.

Outline

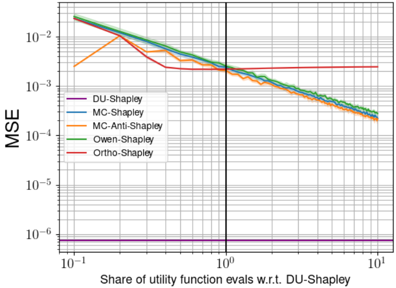

Figure: DU-Shapley vs MC-based approximations on synthetic datasets. Dataset size drawn from the Uniform distribution $\mathrm{U}(\{10, \ldots, 10^3\})$ and variance per player from $\mathrm{U}(\{10^{-3}, \ldots, 10\})$.

Outline

Conclusions & Further work

Conclusions

- We have an efficient Shapley value approximation (linear instead of exponential)

- We have theoretical guarantees for our method

- Our method outperforms state of the art Monte Carlo approximation schemes

Conclusions

- We have an efficient Shapley value approximation (linear instead of exponential)

- We have theoretical guarantees for our method

- Our method outperforms state of the art Monte Carlo approximation schemes

Further work

- Extend the method to more general heterogeneous settings

- Design mechanism to incentivise the data-sharing

Thanks