Probably Approximate Shapley Fairness with Applications in Machine Learning

Zijian Zhou[1*] Xinyi Xu[12*], Rachael Hwee Ling Sim[1], Chuan Sheng Foo[2], Kian Hsiang Low[1]

1 Department of Computer Science, National University of Singapore, Singapore
2 Institute for Infocomm Research, A*STAR, Singapore
* Equal Contribution

AAAI
Association for the Advancement of Artificial Intelligence

Outline

- **Background & Motivation**
- Formulation
- Algorithm
- Experiments

Shapley Value (SV)

Excepted marginal contribution of a player in a group.

Utility function

Marginal Contribution of i in permutation $\pi \in \Pi$

$$\phi_i(N, v) = 1/(n!) \sum_{\pi \in \Pi} \sigma_i(\pi), \text{where } \sigma_i(\pi) = v(P_i^\pi \cup \{i\}) - v(P_i^\pi)$$

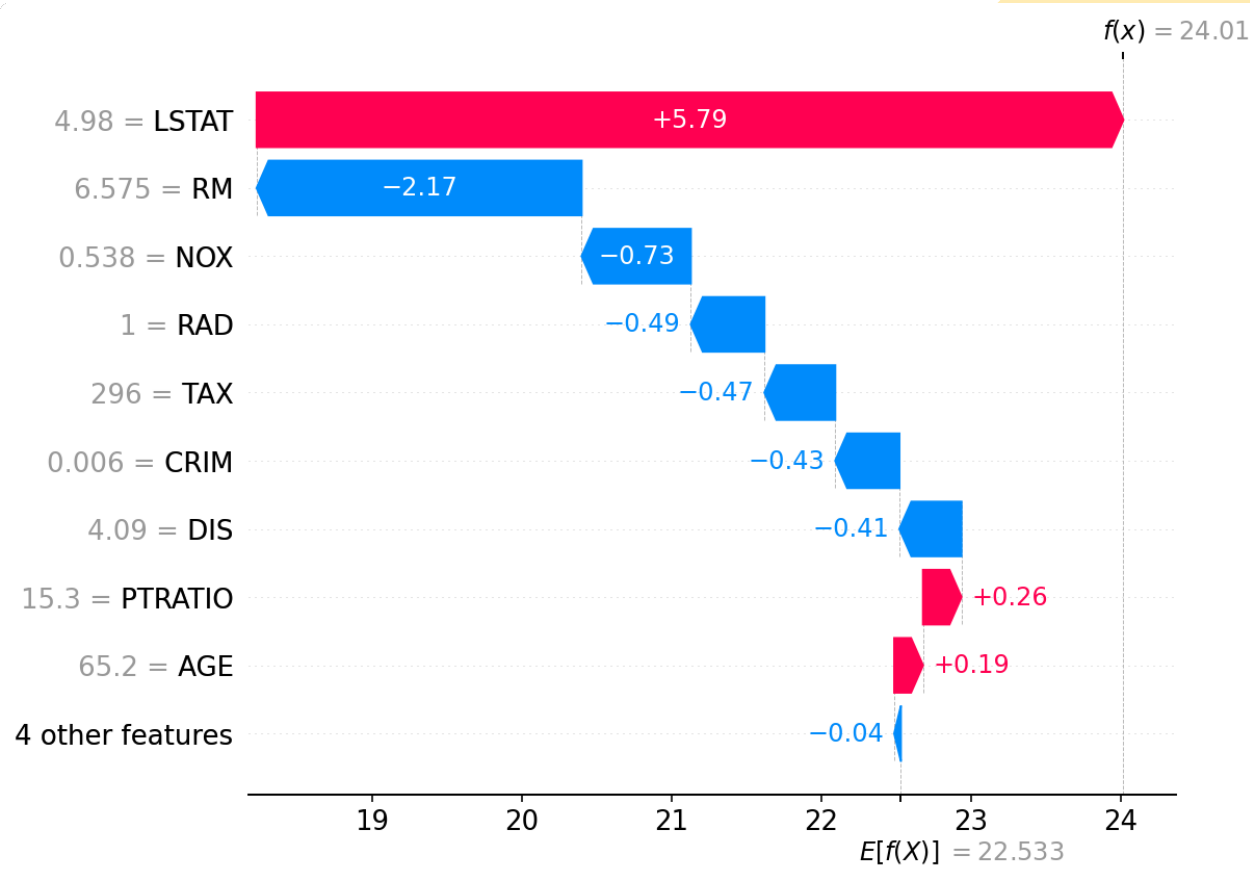
Number of players

SV in Machine Learning

- Feature attribution
- Federated learning
- Data Valuation
- etc.

E.g., feature attribution with SHAP[12]

SV of each feature explains its importance in model performance



1 Lundberg et al., A Unified Approach to Interpreting Model Predictions, NeurIPS 2017
2 Image source: https://github.com/slundberg/shap

Why SV?

Because SV is **FAIR**

SV satisfies a set of **fairness axioms:**

- **Nullity**: if a player i contributes nothing to any subgroup, then $\phi_i = 0$.
- **Symmetry**: if i and j contributes equally to every subgroup, then $\phi_i = \phi_j$.
- **(Strict) Desirability**: if i contributes more or equal than j to every subgroup (and with at least one strictly greater contribution), then $\phi_i > \phi_j$.

SV Computation

However, computing SV is **expensive**

$$\phi_i(N, v) = 1/(n!) \sum_{\pi \in \Pi} \sigma_i(\pi)$$

Iterate through all $|\Pi| = N!$ different permutations

Exponential in terms of N

Feature attribution: N is usually in the range $\mathbf{100 - 1000}$
Federated learning: N is usually **at least 10**
Data valuation: N is usually **at least 1000**

to be applied in practice

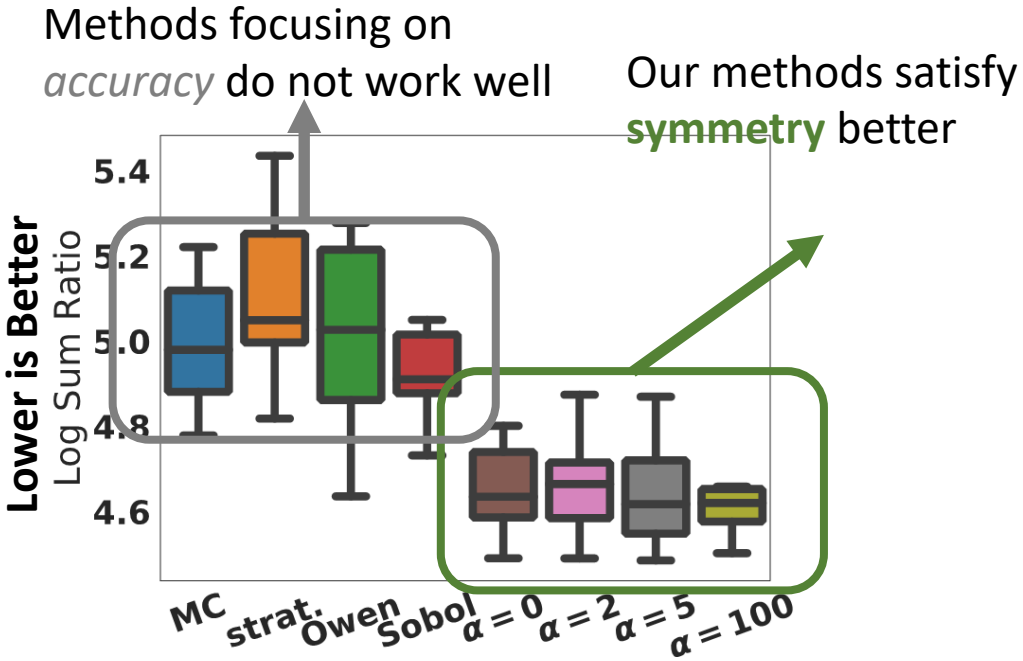
Exact Computation  Estimation 

SV Estimation

- Existing methods try to estimate SV with high *accuracy*

- However, they neglected the important aspect of **fairness** in their approximations.

Metric	Objective
Accuracy	*Mean Squared Error:* $\sum(\phi_i - \hat{\phi}_i)^2$
Fairness	**Various Fairness Properties: Nullity, Symmetry, etc..**



Outline

- Background & Motivation
- **Formulation**
- Algorithm
- Experiments

SV Estimation with Fairness Consideration

Exact Shapley Fairness
Axioms for **Exact SV**

Generalized to

Probably Approximate Shapley Fairness
Axioms for **SV Estimates**

Probably Approximately Correct Learning Framework:

- With probability at least $1 - \delta$, an algorithm A outputs a hypothesis h with error $\leq \epsilon$

Probably Approximate Shapley Fairness:

- With probability at least $1 - \delta$, the Shapley estimate satisfies the fairness properties with error $\leq \epsilon$

Probabilistic Version of Shapley Fairness

- **Nullity**: (the conditional event that) for every $i \in N, |\varphi_i| \leq \epsilon_2$ given that $\phi_i = 0$.
- **Symmetry**: (the conditional event that) for all $i \neq j \in N, \left|\varphi_i - \varphi_j\right| \leq (\epsilon_1|\phi_i| + \epsilon_2) + \left(\epsilon_1\left|\phi_j\right| + \epsilon_2\right)$ given that $\phi_i = \phi_j$.
- **Desirability**: (the conditional event that) $\varphi_i - \varphi_j \geq -(\epsilon_1|\phi_i| + \epsilon_2) - \left(\epsilon_1\left|\phi_j\right| + \epsilon_2\right)$ given that $(\exists B \subseteq N\backslash\{i,j\}, v(B \cup \{i\}) > v(B \cup \{j\})) \wedge (\forall C \subseteq N\backslash\{i,j\}, v(C \cup \{i\}) > v(C \cup \{j\}))$.

Here, φ_i denotes the estimate of ϕ_i. ϵ_1 and ϵ_2 are parameters regulating the tolerance for estimation error.

Objective of SV estimation: Fix a level of error (ϵ_1, ϵ_2), satisfy Probably Approximate Shapley Fairness Axioms with at least probability $1 - \delta$.

↓

Fairness Guarantee

Measuring the level of fairness of SV estimates

Exploit the concept of Signal-to-Noise ratio

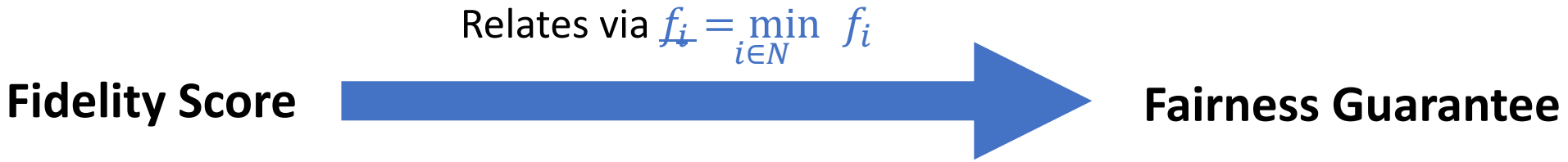
Fidelity Score (FS): the fidelity score of an SV estimate φ_i is defined as

$$f_i = \frac{\left(\left|\phi_i + \frac{\epsilon_2}{\epsilon_1}\right|\right)^2}{\mathrm{Var}(\varphi_i)}$$

Signal with error tolerance

Noise

Fidelity Score and Fairness Guarantee



Key Result (**Proposition 1** in our paper):

- If φ_i's are independent, then with probability at least $\left(1 - \frac{1}{\epsilon_1^2 f_i}\right)^n$, the three probably approximate fairness properties are satisfied within error ϵ_1.
- Otherwise, if φ_i's are dependent, the probability is at least $1 - \frac{n}{\epsilon_1^2 f_i}$.

KEY TAKEAWAY: higher f_i means better fairness guarantee.

Outline

- Background & Motivation
- Formulation
- **Algorithm**
- Experiments

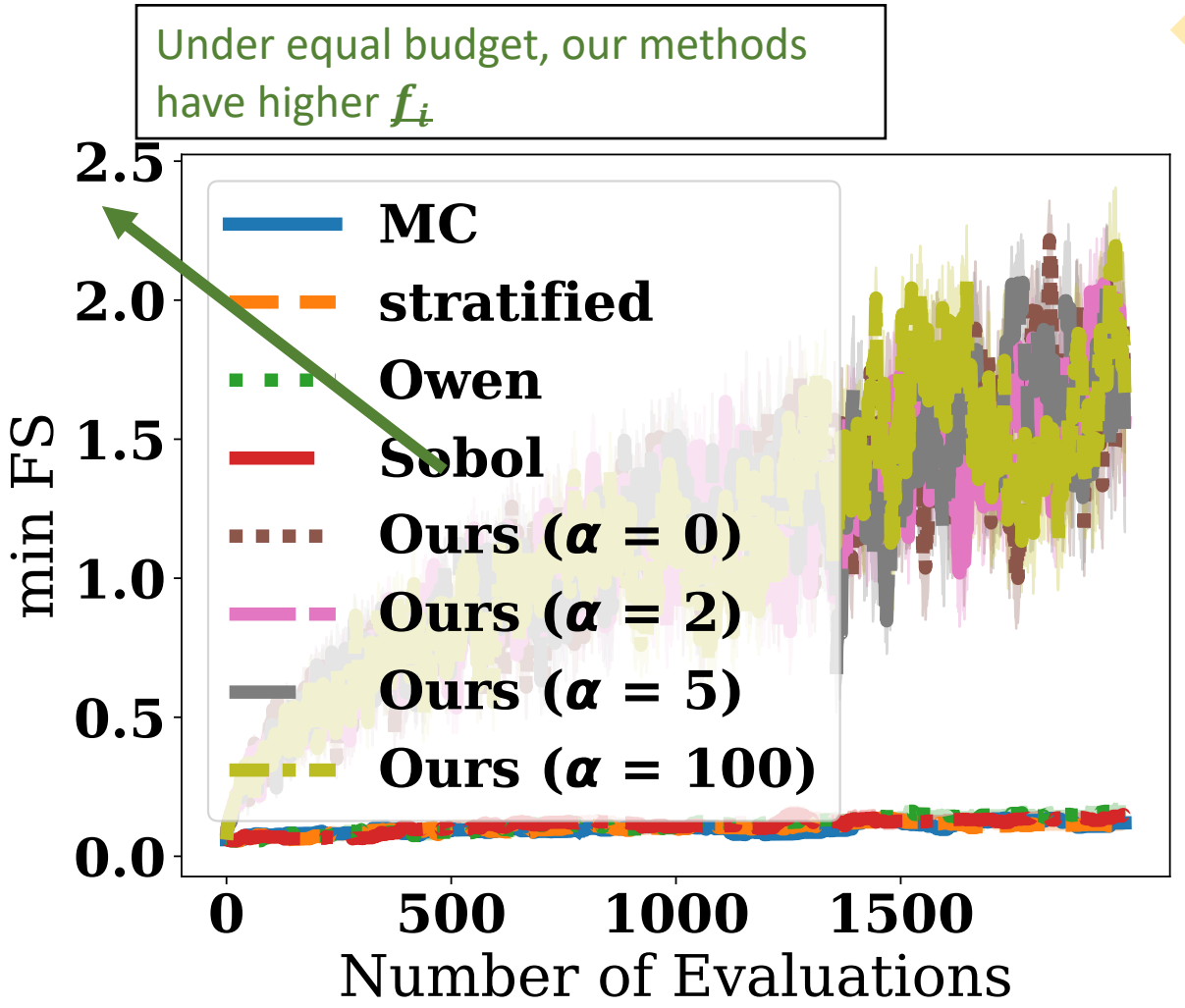
Estimate SV with Fairness Guarantee

Main Idea: **Optimize f_i** to improve fairness guarantee.

Step 1. **Greedily** evaluate the player with the smallest fidelity score.

Step 2. Use **importance sampling** to further reduce estimation noise.

Key Result (**Proposition 3**): with equal budget, Step 1 & Step 2 combined produces higher f_i than classic Monte Carlo method (abbr. as MC on the right).



Outline

- Background & Motivation
- Formulation
- Algorithm
- **Experiments**

Experiments

Experiment conducted

- under various **ML application scenarios**
- with **different datasets**
- to verify the **three probably approximate fairness axioms**

A1: Nullity Property with Agent Valuation[1]

- SV represents the contribution of each agent in FL/CML settings
- Dataset: Hotel reviews with 10 agents, each with a subset of 1000 samples
- Metric: Error of SV with small absolute values after standardization; Lower is better.

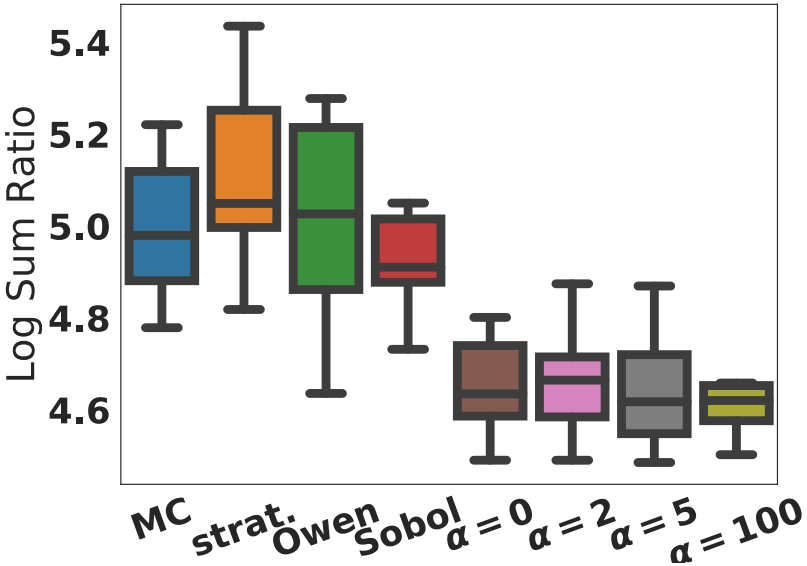
Method	$\sum_{i:\lvert\phi_i\rvert\leq 0.01} \lvert\varphi_i - \phi_i\rvert$, where $\phi^T 1 = \varphi^T 1 = 1$
MC	4.40
Owen Sampling	3.70
Sobol Sampling	8.54
Stratified Sampling	4.10
KernelSHAP	48.9
Ours	**1.40**

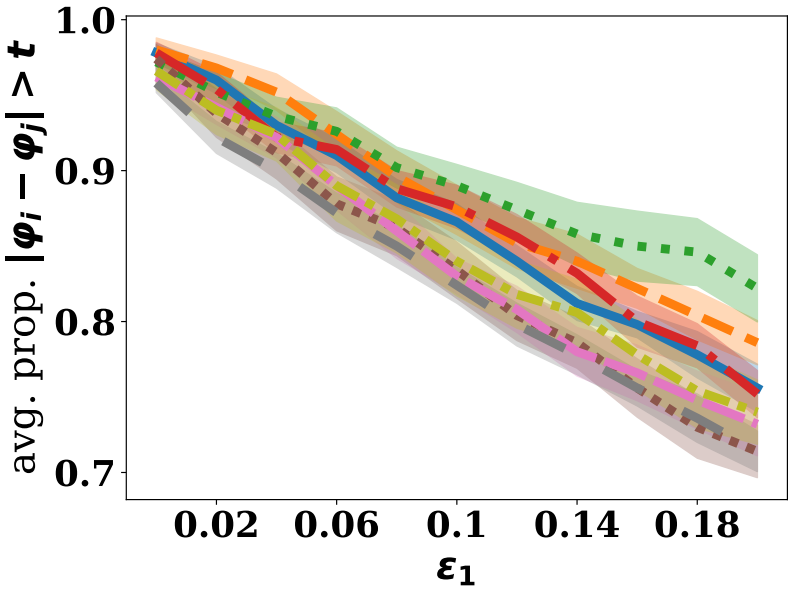
1 Xu et al, Gradient driven rewards to guarantee fairness, NeurIPS 2021

A2: Symmetry Property with Datapoint Valuation[1]

- SV represents the value of each datapoint in the dataset
- Dataset: Breast Cancer (with each data point **duplicated**)
- Metric: For each duplication pair (i, i'), calculate:

1. log sum ratio $\log \sum_{i,i'} \max(\frac{\varphi_i}{\varphi_{i'}}, \frac{\varphi_{i'}}{\varphi_i})$
2. the average proportion where the absolute error exceeds a threshold. For both, lower is better





1 Ghorbani et al, Equitable Valuation of Data for Machine Learning, ICML 2019

A3: Desirability Property with Feature Attribution[1]

- SV represents the importance of each feature
- Dataset: Wine dataset with 7 features extracted by Principal Component Analysis
- Metric: Number of inversions where relative order of SVs change in estimates; Lower is better

Method	$N_{\text{inv}} = \sum_{i \neq j} \mathbf{1}_{\phi_i > \phi_j \cap \varphi_i < \varphi_j} + \mathbf{1}_{\phi_i < \phi_j \cap \varphi_i > \varphi_j}$
MC	0.4
Owen Sampling	1.2
Sobol Sampling	2.4
Stratified Sampling	0
KernelSHAP	3.6
Ours	**0**

1 Lundberg et al., A Unified Approach to Interpreting Model Predictions, NeurIPS 2017

Discussions and Future Work

Summary

- We identified an important yet overlooked **problem with applying SV** in a variety of ML applications
- We formulate a framework for **analyzing the level of Shapley fairness** of SV estimates
- Under the framework, we introduce **a simple yet effective algorithm** to estimate SV with theoretical Shapley fairness guarantee

Future Work

- We focus on designing an algorithm for **unbiased estimates** of SV. Are there any biased estimates that can achieve better probably approximate fairness?

Thank you!