Probably Approximate Shapley Fairness with Applications in Machine Learning

Zijian Zhou[1*], Xinyi Xu[12*], Rachael Hwee Ling Sim[1], Chuan Sheng Foo[23], Bryan Kian Hsiang Low[1]

[1]Department of Computer Science, National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Centre for Frontier AI Research, A*STAR, Singapore
{zhou_zijian, xinyi.xu, rachael.sim}@u.nus.edu, foo_chuan_sheng@i2r.a-star.edu.sg, lowkh@comp.nus.edu.sg

Abstract

The *Shapley value* (SV) is adopted in various scenarios in machine learning (ML), including data valuation, agent valuation, and feature attribution, as it satisfies their fairness requirements. However, as exact SVs are infeasible to compute in practice, SV estimates are approximated instead. This approximation step raises an important question: *do the SV estimates preserve the fairness guarantees of exact SVs?* We observe that the fairness guarantees of exact SVs are too restrictive for SV estimates. Thus, we generalise Shapley fairness to *probably approximate Shapley fairness* and propose fidelity score, a metric to measure the variation of SV estimates, that determines how probable the fairness guarantees hold. Our last theoretical contribution is a novel *greedy active estimation* (GAE) algorithm that will maximise the lowest fidelity score and achieve a better fairness guarantee than the *de facto* Monte-Carlo estimation. We empirically verify GAE outperforms several existing methods in guaranteeing fairness while remaining competitive in estimation accuracy in various ML scenarios using real-world datasets.

1 Introduction

The *Shapley value* (SV) is widely used in machine learning (ML), to value and price data (Agarwal, Dahleh, and Sarkar 2019; Ohrimenko, Tople, and Tschiatschek 2019; Ghorbani and Zou 2019; Ghorbani, Kim, and Zou 2020; Xu et al. 2021b; Kwon and Zou 2022; Sim, Xu, and Low 2022; Wu, Shu, and Low 2022), value data contributors in collaborative machine learning (CML) (Sim et al. 2020; Tay et al. 2022; Agussurja, Xu, and Low 2022; Nguyen, Low, and Jaillet 2022) and federated learning (FL) (Song, Tong, and Wei 2019; Wang et al. 2020; Xu et al. 2021a) to decide fair rewards, and value features' effects on model predictions for interpretability (Covert and Lee 2021; Lundberg and Lee 2017). We consider data valuation as our main example. Given a set N of n training examples and a utility function v that maps a set P of training examples to a real-valued utility (e.g., test accuracy of an ML model trained on P), the SV ϕ_i of the i-th example is

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

where π is a permutation of the N training examples and Π denotes the set of all possible permutations. The SV for the i-th example is its average marginal contribution, $\sigma_i(\pi)$, across all permutations. The marginal contribution $\sigma_i(\pi)$ measures the improvement in utility (e.g., test accuracy) when the i-th example is added to the predecessor set P_i^π containing all training examples j preceding i in π.

The wide adoption of SV is often justified through its fairness axiomatic properties (recalled in Sec. 2). In data valuation, SV is desirable as it ensures that any two training example that improves the ML model performance equally when added to any data subset (i.e., all marginal contributions are equal) are assigned the same value (known as *symmetry*). However, a key downside of SV is that the exact calculation of ϕ_i in Equ. (1) has exponential time complexity and is intractable when valuing more than hundreds of training examples (Ghorbani and Zou 2019; Jia et al. 2019). Existing works address this downside by viewing the SV definition in Equ. (1) as an expectation over the uniform distribution U over Π, $\phi_i = \mathbb{E}_{\pi \sim U}[\sigma_i(\pi)]$, and applying Monte Carlo (MC) approximation (Castro, Gómez, and Tejada 2009) with m_i randomly sampled permutations, thus $\phi_i \approx \varphi_i \coloneqq 1/m_i \sum_{t=1}^{m_i} \sigma_i(\pi^t), \pi^t \sim U$ (Ghorbani and Zou 2019; Jia et al. 2019; Song, Tong, and Wei 2019).

However, this approximation creates an important issue — *(i) do the fairness axioms that justify the use of Shaplay value still hold after approximation (Sundararajan and Najmi 2020; Rozemberczki et al. 2022)?* The answer is unfortunately no as we empirically demonstrate that the symmetry axiom does not hold after approximations in Fig. 2. For two *identical* training examples, their (approximated) SVs used in data pricing are not guaranteed to be equal. We address this unresolved important issue by proposing the notion of *probably approximate Shapley fairness* for SV estimates φ_i, for every $i \in N$. As the original fairness axioms are too restrictive, in Sec. 3, we relax the fairness axioms to *approximate* fairness and consider how they can be satisfied with high *probability*. We introduce a *fidelity score* (FS) to measure the approximation quality of φ_i w.r.t. ϕ_i for each example i and provide a fairness guarantee dependent on the worst/lowest fidelity score across all training examples.

In data valuation, computing the marginal contribution of an $i \in N$ in any sampled permutation is expensive as it involves training model(s). *(ii) How do we achieve probably*

*These authors contributed equally.

approximate Shapley fairness with the lowest budget (number of samples) of marginal contribution evaluations? While it is difficult to achieve the highest approximate fairness possible, we show that we can instead achieve a high fairness *guarantee* (i.e., a lower bound to probably approximate Shapley fairness) via the insight that the budget need not be equally spent on all training examples. For example, if the marginal contribution of example i, $\sigma_i(\pi^t)$ in many sampled permutations are constant, we should instead evaluate that of example j with widely varying $\sigma_j(\pi^t)$ sampled so far. Our method may use a different number of marginal contribution samples, m_i, for each example i and greedily improve the current worst fidelity score across all training examples.

Lastly, to improve the fidelity score, we novelly *use previous samples, i.e., evaluated marginal contribution results, to influence and guide the current sampling of permutations*. In existing MC methods (Owen 1972; Okhrati and Lipani 2020; Mitchell et al. 2022) the sampling distribution that generates π is pre-determined and fixed across iterations. In our work, we use importance sampling to generate π for φ_i as it supports using an alternative proposal sampling distribution, $q_i(\pi)$. For any example i, we constrain permutations π with predecessor set of equal size to have the same probability $q_i(\pi)$. The parameters of the sampling distribution q_i are actively updated across iterations and learnt from past results (i.e., tuples of predecessor set size and marginal contribution) via maximum likelihood estimation or a Bayesian approach. By doing so, we reduce the variance of the estimator φ_i as compared to standard MC sampling, thus improving the fidelity scores efficiently and the overall fairness (guarantee).

Our specific contributions are summarized as follows:

- We propose a *probably approximate Shapley fairness* for SV estimates and exploit an error-aware *fidelity score* to provide a fairness guarantee via a polynomial budget complexity.
- We design greedy selection, which by iteratively prioritising φ_i with lowest FS, can obtain the optimal minimum FS given a fixed total budget m and improve the fairness guarantee (Proposition 1).
- We derive the optimal categorical distribution (intractable) for selecting permutations, and obtain an approximation for active permutation selection. We integrate both greedy and active selection into a novel *greedy active estimation* (GAE) with provably better fairness than MC.
- We empirically verify that GAE outperforms existing methods in guaranteeing fairness while remaining competitive in estimation accuracy in training example and dataset valuations, agent valuation (in CML/FL) and feature attribution with real-world datasets.

2 Preliminaries

Fairness of SV. SV (Equ. (1)) is often adopted (Agarwal, Dahleh, and Sarkar 2019; Sim et al. 2020; Song, Tong, and Wei 2019; Xu et al. 2021a; Sim, Xu, and Low 2022) for guaranteeing fairness by satisfying several axioms (Chalkiadakis, Elkind, and Wooldridge 2011):

F1. Nullity: $(\forall \pi \in \Pi, \sigma_i(\pi) = 0) \implies \phi_i = 0$.

F2. Symmetry: $(\forall C \subseteq N \setminus \{i,j\}, v(C \cup \{i\}) = v(C \cup \{j\})) \implies \phi_i = \phi_j$.

F3. Strict desirability (Bahir et al. 1966): $\forall i \neq j \in N, (\exists B \subseteq N \setminus \{i,j\}, v(B \cup \{i\}) > v(B \cup \{j\})) \wedge (\forall C \subseteq N \setminus \{i,j\}, v(C \cup \{i\}) \geq v(C \cup \{j\})) \implies \phi_i > \phi_j$.

Nullity means if a training example does not result in any performance improvement (marginal contribution is 0 to any permutation), then its value is 0. It ensures offering useless data does not give any reward (Sim et al. 2020)). Symmetry ensures identical values for identical training examples. Strict desirability implies if i gives a larger performance improvement than j in all possible permutations, then i is strictly more valuable than j. We exclude the efficiency axiom as it does not suit ML use-cases (Ghorbani and Zou 2019; Bian et al. 2022; Kwon and Zou 2022),[1] and exclude the linearity (Jia et al. 2019) and monotonicity (Sim et al. 2020) axioms for fairness analysis as we restrict our consideration to one utility function v (Bian et al. 2022). Note that the fairness of exact SV is *binary*: either satisfying all these axioms or not.

Sampling-based estimations. These methods typically extend the MC formulation of $\phi_i \approx \varphi_i := \mathbb{E}_{\pi_t \sim U}[\sigma_i(\pi_t)]$ by changing the sampling distribution of π_t. Importantly, in all these methods, for each sampled permutation π_t, a *single* marginal contribution $\sigma_i(\pi_t)$ is evaluated and used in φ_i. Thereafter, we consistently refer to this single evaluation as expending one budget (i.e., one permutation) and the corresponding marginal contribution $\sigma_i(\pi_t)$ as one sample.

Formally, a sampling-based estimation method estimates ϕ_i via the expectation of the random variable $\sigma_i(\pi)$ (which depends on the permutations randomly sampled according to some distribution q): $\phi_i \approx \mathbb{E}_q[\sigma_i(\pi)]$. Hence, such methods differ from each other in the distribution q as well as the selection of entry $i \in N$ to evaluate in each iteration. The estimates φ_i's can be independent of each other such as MC (Castro, Gómez, and Tejada 2009), and stratified sampling (Maleki et al. 2013; Castro et al. 2017), or dependent such as antithetic Owen method (Owen 1972; Okhrati and Lipani 2020) and Sobol method (Mitchell et al. 2022). We provide theoretical results for both scenarios when the estimates φ_i's are independent and dependent.

3 Generalised Fairness For Shapley Value Estimates

3.1 Fairness Axioms For Shapley Value Estimates

We propose the following re-axiomatisation of fairness for SV estimates (based on axioms F1-F3) using conditional events, by analysing multiplicative and absolute errors.

A1. Nullity: let E_{A_1} be the (conditional) event that for any $i \in N$, conditioned on $\phi_i = 0$, $|\varphi_i| \leq \epsilon_2$.

A2. Symmetry: let E_{A_2} be the (conditional) event that for all $i \neq j \in N$, conditioned on $\phi_i = \phi_j$, then $|\varphi_i - \varphi_j| \leq (\epsilon_1|\phi_i| + \epsilon_2) + (\epsilon_1|\phi_j| + \epsilon_2)$.

[1] Efficiency requires $\sum_i \phi_i = v(N)$ which is difficult to verify in practice for ML (Kwon and Zou 2022); it does not make sense to "distribute" the voting power in feature attribution (interpretation of efficiency) to each i (Bian et al. 2022).

A3. Approximate desirability: let E_{A_3} be the (conditional) event that for all $i \neq j \in N$, conditioned on $(\exists B \subseteq N \setminus \{i,j\}, v(B \cup \{i\}) > v(B \cup \{j\})) \wedge (\forall C \subseteq N \setminus \{i,j\}, v(C \cup \{i\}) \geq v(C \cup \{j\}))$, then $\varphi_i - \varphi_j > -(\epsilon_1|\phi_i| + \epsilon_2) - (\epsilon_1|\phi_j| + \epsilon_2)$.

Thereafter, satisfying A1-A3 refers to the events E_{A_1}-E_{A_3} occurring, respectively. To see A1-A3 generalise the original axioms F1-F3:[2] A1 requires the SV estimate to be small for a true SV with value 0 (as in F1); A2 requires the SV estimates for equal SVs be close (generalised from being equal in F2); A3 requires the ordering of φ_i, φ_i for some i, j from F3 to be preserved up to some error, specifically a multiplicative error ϵ_1 (to account for different $|\phi_i|$) and an absolute error ϵ_2 (to avoid degeneracy from extremely small $|\phi_i|$) where F3 has no such error term. Intuitively, smaller errors ϵ_1, ϵ_2 mean $\boldsymbol{\varphi}$ are "closer" (in fairness) to $\boldsymbol{\phi}$. In addition, the ratio $\xi \equiv \epsilon_2/\epsilon_1$ denotes the tolerance (to be set by user) of relative (multiplicative) vs. absolute errors where a larger ξ implies a higher tolerance for absolute error (i.e., favours A1 over A2 & A3) and *vice versa*. In contrast to existing works that only consider the concentration results of φ_i w.r.t. ϕ_i for each i (Castro, Gómez, and Tejada 2009; Castro et al. 2017; Maleki et al. 2013), we additionally consider the interaction between i, j to define the following:

Definition 1 (**Probably Approximate Shapley Fairness).** For fixed ϵ_1, ϵ_2, and some $\delta \in (0, 1)$ s.t. $\boldsymbol{\varphi}$ satisfy A1-A3 *jointly* w.p. $\geq 1-\delta$, then we call $\boldsymbol{\varphi}$ satisfy $(\epsilon_1, \epsilon_2, \delta)$-Shapley fairness.

In Definition 1, $\boldsymbol{\varphi}$ are probably (w.p. $\geq 1-\delta$) approximately (w.r.t. errors ϵ_1, ϵ_2) Shapley fair. Hence, given the error requirements ϵ_1, ϵ_2, a smaller δ means a better fairness guarantee. In particular, $\boldsymbol{\phi}$ satisfy the optimal $(0, 0, 0)$-Shapley fairness. Despite the appeal, analysing existing estimators w.r.t. Definition 1 is difficult because most do not come with a direct variance result (Zhou et al. 2023). The only expect is the MC method (Castro, Gómez, and Tejada 2009; Maleki et al. 2013) which we analyse in Sec. 4.

3.2 Fairness Guarantee Via The Fidelity Score Of Shapley Value Estimates

Inspired by the concept of *signal-to-noise ratio* (SNR) widely adopted in optics (de Boer et al. 2003) and imaging (Rose 2012), we design a metric for the variation of φ_i, named fidelity score, expressed in Definition 2.

Definition 2 (**Fidelity Score).** The fidelity score (FS) of an (unbiased) SV estimate φ_i for ϕ_i is defined as $f_i \equiv \text{FS}(\varphi_i, \epsilon_1, \epsilon_2) := (|\phi_i| + \epsilon_2/\epsilon_1)^2/\mathbb{V}[\varphi_i]$ where $\mathbb{V}[\varphi_i]$ is the variance of φ_i.[3]

The FS exactly matches the SNR in φ_i when $\epsilon_2 = 0$.[4] A higher f_i implies a more accurate φ_i. f_i is higher when the variance $\mathbb{V}[\varphi_i]$ is small. This occurs when the marginal contributions, $\sigma_i(\pi)$, are close for all permutations $\pi \in \Pi$ or when the number of samples m_i used to compute φ_i is large. As $m_i \to \infty$, $\mathbb{V}[\varphi_i] \to 0$ and $f_i \to \infty$. Additionally, we introduce an error-aware term $\xi := \epsilon_2/\epsilon_1$ in the FS numerator to better capture estimation errors and allow examples with SV of 0 to have different FSs.

Moreover, we empirically verify that f_i is a good reflection of the approximation quality and analyze the impact of ξ. For the former, we compared f_i vs. absolute percentage error (APE) of $\varphi_i := |(\varphi_i - \phi_i)/\phi_i|$.[5] Fig. 1 shows the negative correlation between FS (f_i) and estimation error (APE) (note that we plot $\text{APE}^{-0.5}$). This will justify our proposal to prioritize improving the estimate φ_i with the lowest f_i in Sec 4.1. For the latter, we compare the correlation between FS, f_i and the inverse estimation error, $\text{APE}^{-0.5}$, for different values of ξ in Tab. 1 and find that $\xi =$ 1e-3 leads to a strong positive correlation and is a sweet spot (second best in both settings).

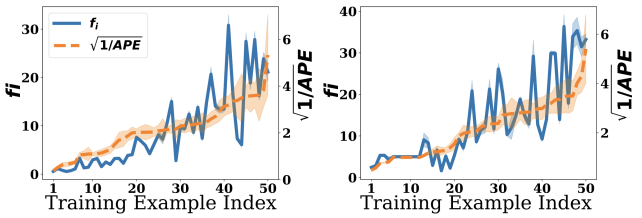


Figure 1: Average (standard error) of $f_i, \text{APE}^{-0.5}$ over 20 trials (sorted in increasing order of $\text{APE}^{-0.5}$ of 50 training examples) with $\xi =$ 1e-3. Left (right) is logistic regression (k-nearest neighbors) using breast cancer (MNIST) dataset.

ξ	breast cancer (logistic)	diabetes (ridge)
1e-5	**6.89e-01** (1.41e-02)	5.49e-01 (2.24e-02)
1e-4	**6.89e-01** (1.41e-02)	5.48e-01 (2.24e-02)
1e-3	6.78e-01 (1.50e-02)	5.53e-01 (2.20e-02)
1e-2	-1.94e-01 (2.59e-02)	**5.67e-01** (2.32e-02)
1e-1	-1.47e-01 (2.60e-03)	5.00e-01 (2.55e-02)

Table 1: Spearman coefficient between f_i and $\text{APE}^{-0.5}$. Average (standard error) over 20 independent trials. Higher is better.

Definition 2 allows us to leverage the Chebyshev's inequality to derive a fairness guarantee, through the minimum FS $\underline{f} := \min_{i \in N} f_i$.

Proposition 1. $\boldsymbol{\varphi}$ satisfy $(\epsilon_1, \epsilon_1\xi, \delta)$-Shapley fairness where $\delta = 1-(1-1/(\epsilon_1^2\underline{f}))^n$ if all φ_i's are independent; otherwise $\delta = n/(\epsilon_1^2\underline{f})$.

Proposition 1 (its proof is in (Zhou et al. 2023)) formalises the effects of the variations in $\boldsymbol{\varphi}$ in satisfying probably ap-

[2] An equivalent formulation for F1-F3 using conditional events are in (Zhou et al. 2023).

[3] Our implementation estimates $|\phi_i|$ and $\mathbb{V}[\varphi_i]$ to obtain f_i.

[4] For a fixed FS, an example i with a larger SV (signal) can contain more noise.

[5] We fit a learner on 50 randomly selected training examples from breast cancer (Street 1995) and MNIST (LeCun et al. 1990) (diabetes (Efron et al. 2004)) datasets and set test accuracy (negative mean squared error) as v using data Shapley (Ghorbani and Zou 2019) with $m_i = 50$. f_i is approximated using sample evaluations. Additional details and results are in (Zhou et al. 2023).

proximate Shapley fairness where a larger minimum variation (i.e., $\underline{f}$) results in larger δ, hence lower probability of $\boldsymbol{\varphi}$ satisfying the fairness axioms. To see whether φ_i's are independent, it is equivalent to checking whether any permutation sampled is used for estimating multiple φ_i's (proof in (Zhou et al. 2023)).

The fidelity score, f_i, is sensitive to the number of sampled permutations and marginal contributions, m_i, as the variance of the SV estimator uses m_i independent samples to produce $\mathbb{V}[\varphi_i] = \mathbb{V}[\sigma_i(\pi)]/m_i$. Therefore, we define an insensitive quantity, the *invariability* of i, r_i, as the FS with only *one* sample of π. Hence, $r_i \propto 1/\mathbb{V}[\sigma_i(\pi)]$ and here $\pi \sim U$. A higher r_i implies that i-th marginal contribution is more invariable across different permutations. The fidelity score is product of the invariability and number of samples, i.e., $f_i = m_i r_i$, used in proving the following corollary.

Corollary 1. Using the notations in Proposition 1, the minimum total budget $m = \sum_{i \in N} m_i$ needed to satisfy $(\epsilon_1, \epsilon_1 \xi, \delta)$-Shapley fairness is at most $\mathcal{O}(n\epsilon_1^{-2}(1 - (1 - \delta)^{1/n})^{-1})$ if φ_i's are independent; and $\mathcal{O}(n^2 \epsilon_1^{-2} \delta^{-1})$ o/w.

The budget complexity is an upper bound in a best/ideal case in the sense that our derivation (in (Zhou et al. 2023)) uses r_i which cannot be observed in practice. While it is not shown to be tight, our $\mathcal{O}(n)$ budget complexity for the independent scenario is linear in terms of the number of training examples, and the $\mathcal{O}(n^2)$ budget complexity for the dependent scenario seems necessary for the $\mathcal{O}(n^2)$ pairwise interactions between φ_i's. Sec. 4 describes an estimation method that runs in the budget complexity upper bound given in Corollary 1.

4 Fairness via Greedy Active Estimation

To achieve probably approximate Shapley fairness with the lowest budget (number of samples) of marginal contribution evaluations, we propose a novel greedy active estimation (GAE) consisting of two core components - *greedy selection* and *importance sampling*. The first component efficiently split the training budget across examples while the second component influence and guide the sampling of permutations and reduce the variance $\mathbb{V}[\varphi_i]$ for each example i. In this section, we outline these components and show that they improve the minimum fidelity scores, $\underline{f}$. The details and full pseudo-code of the algorithm is given in (Zhou et al. 2023).

4.1 Greedy Selection Using Pigou-Dalton

From Proposition 1, we can observe that a larger $\underline{f}$ decreases the probability of unfairness. Hence, to efficiently achieve probably approximate fairness, we should maximize $\underline{f}$ by improving the FS of the training example with the current lowest f_i. Our greedy method is outlined in Proposition 2 and formally proven in (Zhou et al. 2023).

Proposition 2 (*Informal*)**.** Given the constraint of evaluating a total of m marginal contributions for all $j \in N$ to its predecessor set when the permutations are sampled from a fixed distribution q (e.g., the uniform distribution). Then, the minimum FS $\underline{f}$ is maximised by (iteratively) greedily selecting and evaluating a marginal contribution of $i = \operatorname{argmin}_{j \in N} f_j$, until m is exhausted.

One direct implication is that greedy selection outperforms equally allocating the budget m among all N (i.e., $m_i = m/n$), which is used by existing methods (Sec. 2). Greedy selection will use a lower budget m_i on a training example i with higher invariability r_i (lower variance in marginal contribution across permutations) to meet the same threshold $\underline{f}$. The budget will be mainly devoted to training examples with lower invariability and higher variance instead.

Moreover, improving $\underline{f}$ across all $i \in N$ is in line with the Pigou-Dalton principle (PDP) (Dalton 1920): Suppose we have two divisions of the budget that result in two sets of FSs denoted by $\boldsymbol{f}, \boldsymbol{f}' \in \mathbb{R}^n$ respectively. PDP prefers $\boldsymbol{f}$ to $\boldsymbol{f}'$ if $\exists i, j \in N$ s.t., (a) $\forall k \in N \setminus \{i, j\}, f_k = f'_k$ and (b) $f_i + f_j = f'_i + f'_j$ and (c) $|f_i - f_j| < |f'_i - f'_j|$. PDP favors a division of the budget that leads to more equitable distribution of FSs. For SV estimation, it means we are approximately equally sure about all the estimates of the training examples, which can improve the effectiveness of identifying valuable training examples for active learning (Ghorbani, Zou, and Esteva 2021) (Fig. 3) or the potentially noisy/adversarial ones (Ghorbani and Zou 2019) (see results in (Zhou et al. 2023)). Theoretically, an inequitable distribution of FSs with some training examples with significantly lower f_i will have a worse fairness guarantee (Proposition 1). We show that greedy selection satisfies PDP in (Zhou et al. 2023).

Remark. Although greedy selection maximizes the minimums FS, $\underline{f}$, it is not guaranteed to achieve approximate Shapley fairness with the highest probability as the probability bound in Proposition 1 is not tight (e.g., due to the application of union bound in derivation). Nevertheless, in Sec. 5, we empirically demonstrate that greedy selection indeed outperforms other existing methods in achieving probably approximate fairness. It is an appealing future direction to further improve the analysis and provide a tight bound for Proposition 1.

4.2 Active Permutation Selection Via Importance Sampling

To improve the fidelity scores in Definition 2, our GAE method uses importance sampling to reduce $\mathbb{V}[\varphi_i]$ for every training example i by setting $\varphi_i := 1/m_i \sum_{t=1}^{m_i} \sigma_i(\pi^t)/(q_i(\pi^t)n!), \pi^t \sim q_i$. Here, q_i is our proposal distribution over set of all permutations Π that assigns probability $q_i(\pi)$ to permutation π. Following existing works (Castro et al. 2017; Ghorbani, Kim, and Zou 2020), we encode the assumption that any permutation, π, with the same cardinality for the predecessor set of i, P_i^π, should be assigned equal sampling probability. Hence $q_i(\pi) \propto q'_i(|P_i^\pi|)$ where the function q'_i maps the predecessor's cardinality to the sampling probability. We derive the optimal (but intractable) distribution $q_i^*(\pi) \propto q_i'^*(|P_i^\pi|) \propto (\mathbb{E}_{\pi \sim U(|P_i^\pi|)}[\sigma(\pi)^2])^{0.5}$ (proof in (Zhou et al. 2023)) and approximate it with a "learnable" q'_i. Specifically, we use a categorical distribution over the support $\{0, \ldots, n-1\}$ as q'_i and learn its parameters through maximum likelihood estimation (MLE) on tuples of predecessor set size and marginal contribution, with bootstrapping (i.e., sampling a small amount of permutations using MC, detailed in (Zhou et al. 2023)).

This active selection leads to both theoretical (Proposition 3) and empirical improvements (see Sec. 5), whilst ensuring $\mathbb{E}[\varphi_i] = \phi_i$ (Zhou et al. 2023).

Proposition 3. For a fixed budget m, denote the minimum FS achieved by greedy selection and active importance sampling, greedy selection only (with uniform sampling), and uniform MC sampling as $\underline{f}_{\text{GAE}}, \underline{f}_{\text{greedy}}$ and $\underline{f}_{\text{MC}}$, respectively. Then, GAE outperforms the other methods as

$$\underline{f}_{\text{GAE}} \geq \underline{f}_{\text{greedy}} \geq \underline{f}_{\text{MC}} .$$

Furthermore, the minimum FSs $(\underline{f}_{\text{GAE}}, \underline{f}_{\text{greedy}})$ are equal iff (a) $\forall i \in N, \mathbb{V}_{\pi\sim q_i^*}[\sigma_i(\pi)/(q_i^*(\pi)n!)] = \mathbb{V}[\sigma_i(\pi)]$ and the minimum FSs $(\underline{f}_{\text{greedy}}, \underline{f}_{\text{MC}})$ are equal iff (b) every $i \in N$ has the same invariability w.r.t. q_i^*, $r_{i,q_i^*} := (|\phi_i| + \xi)^2/\mathbb{V}_{\pi\sim q_i^*}[\sigma_i(\pi)/(q_i^*(\pi)n!)]$.[6]

The proof is in (Zhou et al. 2023). In practice, equality conditions (a) and (b) are unlikely to hold. (a) only holds when our cardinality assumption (i-th marginal contribution to predecessor set of the same cardinality are similar) is wrong and unhelpful. A necessary but unrealistic condition for (b) is that for two data points i, j with the same SV, their marginal contributions, i.e., the set $\{\sigma_i(\pi)|\pi \in \Pi\}$, must be equal.

Regularising importance sampling with uniform prior. Proposition 3 requires the cardinality assumption so that the importance sampling approach (using q_i' which corresponds to a q_i) is effective by ensuring the derived q_i is close to the theoretical optimal q_i^*. However, in practice, there are situations where using the uniform distribution U performs better than q_i obtained via learning q_i' using the cardinality assumption. First, if the marginal contributions on the *same* cardinality vary significantly (i.e., the cardinality assumption does not hold), then the approximation using q_i is inaccurate. Second, if the marginal contributions over *different* cardinalities vary minimally, then importance sampling has little benefit as the marginal contributions are approximately equal. Interestingly, in both cases, using U (treating all cardinalities uniformly) is the mitigation because it avoids using the incorrect inductive bias (i.e., the cardinality assumption).

Therefore, to incorporate U, we regularise the learning of q_i' with a uniform Dirichlet prior. Specifically, from the MLE parameters $\boldsymbol{w} \in \triangle(n)$ of $q_i'^*, i \in N$ obtained via bootstrapping,[7] and a uniform/flat Dirichet prior parameterised by $(\alpha+1)\mathbf{1}_n, \alpha \geq 0$, we can obtain the maximum a posteriori (MAP) estimate as $n\boldsymbol{w} + (\alpha+1)\mathbf{1}_n$ (more details and derivations in (Zhou et al. 2023)). With this, we unify the frequentist approach of learning $q_i'^*$'s parameters via purely MLE and the Bayesian approach of incorporating both likelihood and prior belief with α controlling how strong our prior belief is. Specifically, when $\alpha = 0$, q_i' reduces to MLE and as $\alpha \to \infty$, q_i' tends to the uniform distribution over cardinality (due to the uniform Dirichlet prior) and thus the corresponding q_i tends to U. In other words, if the cardinality assumption is satisfied (not satisfied), we expect the learned q_i' with a small (large) α to perform better (Sec. 5).

5 Experiments

We empirically verify that our proposed method can effectively mitigate the adverse situation described in introduction — the violation of the original symmetry axiom and its negative consequences (e.g., identical data are valued/priced very differently). We further compare GAE's performance w.r.t. other axioms and PDP, and in different scenarios with real-world datasets against existing methods. We provide the complete details on setting, implementation and experimental results in (Zhou et al. 2023) and the code is available at https://github.com/BobbyZhouZijian/ProbablyApproximateShapleyFairness.

Specific problem scenarios and comparison baselines. As mentioned in Sec. 2, our method is general, so we empirically investigate several specific problem scenarios in ML: **P1.** Data point valuation quantifies the relative effects of each training example in improving the learning performance (to remove noisy training examples or actively collect more valuable training examples) (Bian et al. 2022; Ghorbani, Kim, and Zou 2020; Ghorbani and Zou 2019; Jia et al. 2019; Jia et al. 2019; Kwon and Zou 2022). **P2.** Dataset valuation aims to provide value of a dataset among several datasets (e.g., in a data marketplace) (Agarwal, Dahleh, and Sarkar 2019; Ohrimenko, Tople, and Tschiatschek 2019; Xu et al. 2021b). **P3.** Agent valuation in the CML/FL setting determines the contributions of the agents to design their compensations (Sim et al. 2020; Song, Tong, and Wei 2019; Wang et al. 2020; Xu et al. 2021a). **P4**. Feature attribution studies the relative importance of features in a model's predictions (Covert and Lee 2021; Lundberg and Lee 2017). We investigate **P1.** in detail in Sec. 5.1 and **P2. P3.** and **P4.** together in Sec. 5.2. We compare with the following estimation methods (as baselines): MC (Castro, Gómez, and Tejada 2009), stratified sampling (Castro et al. 2017; Maleki et al. 2013), multi-linear extension (Owen) (Okhrati and Lipani 2020; Owen 1972), Sobol sequences (Mitchell et al. 2022) and improved KernelSHAP (kernel) (Covert and Lee 2021).[8]

5.1 Investigating A1-A3 And PDP Within P1.

Settings. We fit classifiers (e.g., logistic regression) on different datasets, use test accuracy as v (Ghorbani and Zou 2019; Jia et al. 2019), and adopt Data Shapley (Ghorbani and Zou 2019, Proposition 2.1) as the SV definition. We randomly select 50 training examples from a dataset and duplicate each once (i.e., a total of $n = 100$ training examples). Following Sec. 3, we set $\xi =$ 1e-3. For bootstrapping (included for all baselines), we uniformly randomly select 20 permutations and evaluate the marginal contributions for each i. We set a budget $m = 2000$ for each baseline. As the true $\boldsymbol{\phi}$ are

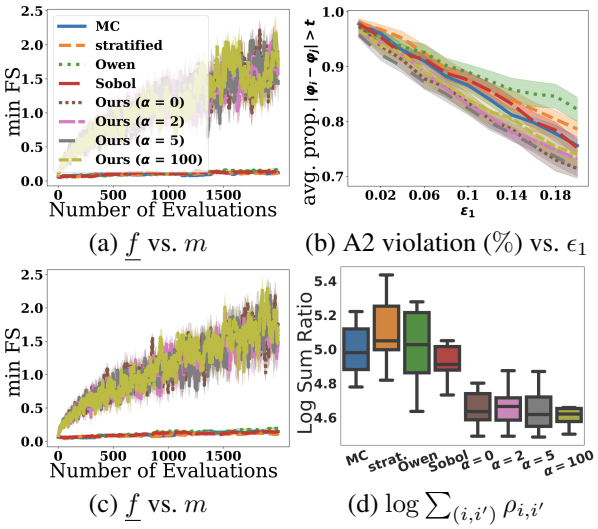
[6] As before, the invariability is the FS when using only *one* sample, but the definition is updated to match the redefined φ_i.

[7] $\triangle(n)$ denotes the probability simplex of dimension $n-1$ and $\boldsymbol{w}$ is derived in (Zhou et al. 2023).

[8] We follow (Covert and Lee 2021) as it provides an unbiased estimator where the original estimator (Lundberg and Lee 2017) is only provably consistent and empirically shown to be less efficient (Covert and Lee 2021).

intractable for $n = 100$, $\boldsymbol{\phi}$ is approximated via MC with significantly larger budget (200 bootstrapping permutations and $m = 30000$, averaged over 10 independent trials) as ground truth (in order to evaluate estimation errors) (Jia et al. 2019).

Effect of $\underline{f}$ on symmetry A2. Proposition 1 implies a higher $\underline{f}$ leads to a better fairness guarantee. We specifically investigate the effect of $\underline{f}$ in mitigating mis-estimations of *identical* training examples and directly verify A2. We consider three evaluation metrics: lowest FS (i.e., $\underline{f}$); the proportion of duplicate pairs i, i' with estimates having a deviation larger than a threshold t, i.e., $|\varphi_i - \varphi_{i'}| > t = \epsilon_1|\phi_i| + \xi\epsilon_1$ (as in A2); and the (log of) sum of the deviation ratio $\rho_{i,i'} := \max(\varphi_i/\varphi_{i'}, \varphi_{i'}/\varphi_i)$ over all i, i' pairs. In Fig. 2a,c, the $\underline{f}$ of our methods increases (improves) as the number of samples, m, increases. However, the $\underline{f}$ of all baseline methods are stagnated close to 0. Thus, as expected, our method significantly outperforms the baselines in obtaining a high $\underline{f}$. As predicted by Proposition 1, this results in a lower probability and extent of fairness. As compared to baseline methods, our methods consistently have a lower proportion of identical examples that do not satisfy symmetry (A2) (Fig. 2b) and smaller deviation ratio of the estimated SV (Fig. 2d).



(a) $\underline{f}$ vs. m (b) A2 violation (%) vs. ϵ_1

(c) $\underline{f}$ vs. m (d) $\log \sum_{(i,i')} \rho_{i,i'}$

Figure 2: (a-b) and (c-d) plot results from using logistic regression on the synthetic Gaussian (Kwon and Zou 2022) and Covertype (Blackard 1998) datasets, respectively, using various baseline methods and ours. In (a,c), higher $\underline{f}$ is better while in (b,d), lower values are better. The intervals show the standard error of 10 independent trials.

Verifying nullity A1 and Pigou Dalton Principle. In practice, A1 is rarely applicable (i.e., $\forall\pi, \sigma_i(\pi) = 0$), so we instead investigate a more likely scenario: $|\phi_i|$ is very small for some i, because the training example i has a minimal impact during training (e.g., i is redundant). We randomly draw 20 training examples from the breast cancer dataset (Street 1995) to fit a support vector machine classifier (SVC). To verify A1, we standardize $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ (i.e., $\boldsymbol{\phi}^\top \mathbf{1}_n = \boldsymbol{\varphi}^\top \mathbf{1}_n = 1$) and calculate $\epsilon_{\text{abs}} := \sum_{i:|\phi_i|\leq 0.01} |\varphi_i - \phi_i|$. As PDP is difficult to verify directly but it satisfies the Nash social welfare (NSW) (de Clippel 2010; Kaneko and Nakamura 1979), we use the (negative log of) NSW on standardized $\boldsymbol{f}$ (i.e., $\boldsymbol{f}^\top \mathbf{1}_n = n$) NL NSW $:= -\log \prod_{i\in N} f_i$ (lower indicates PDP is better satisfied). Tab. 2 shows our method obtains lowest estimation errors on (nearly) null training examples and satisfies PDP the best.

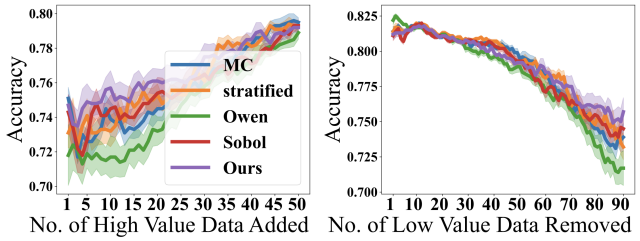


Figure 3: Accuracy of logistic regression adding (removing) training examples generated from Gaussian distribution with highest (lowest) φ_i in left (right). Bootstrapping with 20 permutations and $m = 2000$, $\alpha = 5$ for our method. Average (standard errors in shaded color) over 10 independent trials. KernelSHAP does not give clear trends in both plots because its estimates are not very accurate. Therefore we omit it here. A more detailed plot which involves KernelSHAP and other variants of our methods is provided in (Zhou et al. 2023).

baselines	ϵ_{abs}	NL NSW
MC	2.12e-02 (2.75e-03)	14.6 (5.12e-01)
Owen	6.33e-02 (4.04e-03)	18.2 (7.10e-01)
Sobol	6.28e-02 (5.64e-03)	11.6 (6.24e-01)
stratified	2.89e-02 (5.84e-03)	14.0 (7.46e-01)
kernel	6.44e-01 (1.15e-01)	21.4 (1.48)
Ours ($\alpha = 0$)	1.72e-02 (3.50e-03)	**2.24** (7.40e-01)
Ours ($\alpha = 2$)	1.61e-02 (4.65e-03)	2.57 (5.28e-01)
Ours ($\alpha = 5$)	**1.42e-02** (4.00e-03)	2.55 (5.51e-01)
Ours ($\alpha = 100$)	2.13e-02 (5.95e-03)	3.48 (5.61e-01)

Table 2: Average (standard errors) of error and NL NSW over 10 independent trials.

Verifying approximate desirability A3. For A3, we verify whether the valuable training examples have high φ_i by adding (removing) training examples according to highest (lowest) φ_i (Ghorbani and Zou 2019; Kwon and Zou 2022) (Fig. 3) and via noisy label detections (Jia et al. 2019) (see results in (Zhou et al. 2023)). Fig. 3 left (right) shows our method is effective in identifying the most (least) valuable training examples to add (remove).

5.2 Generalising To Other Scenarios

We examine the estimation accuracy, A3, and PDP within **P2.**, **P3.** and **P4.**.[9] For **P2.** we adopt robust volume SV (Xu

[9]We exclude **P1.** and **P2.** because they are less likely to be applicable in these scenarios.

baselines	MAPE	MSE	N_{inv}	ϵ_{inv}	NL NSW
MC	3.87e-02 (7.9e-03)	2.70e-03 (7.9e-04)	3.60 (1.33)	4.58 (0.82)	1.72e-02 (4.6e-03)
Owen	3.06e-02 (6.7e-03)	1.60e-03 (5.2e-04)	4.00 (1.41)	3.50 (0.76)	1.31e-02 (3.6e-03)
Sobol	6.75e-02 (3.4e-03)	9.62e-03 (1.5e-03)	4.80 (1.20)	7.97 (0.56)	7.46e-02 (1.3e-02)
stratified	4.46e-02 (8.3e-03)	3.30e-03 (8.3e-04)	4.40 (1.17)	5.17 (0.87)	1.72e-02 (5.9e-03)
kernel	0.10 (2.0e-02)	1.37e-02 (3.7e-03)	8.80 (2.15)	1.09e+01 (2.02)	3.64 (0.46)
Ours ($\alpha = 0$)	0.10 (1.6e-02)	2.15e-02 (8.0e-03)	1.08e+01 (2.24)	1.18e+01 (1.85)	2.60 (0.88)
Ours ($\alpha = 2$)	2.30e-02 (2.5e-03)	7.60e-04 (1.7e-04)	2.80 (1.02)	2.50 (0.29)	**3.40e-04** (9.0e-05)
Ours ($\alpha = 5$)	**2.14e-02** (3.1e-03)	**6.80e-04** (1.7e-04)	**1.20** (0.49)	**2.34** (0.31)	9.90e-04 (9.0e-05)
Ours ($\alpha = 100$)	2.40e-02 (2.9e-03)	9.90e-04 (2.8e-04)	2.40 (0.75)	2.77 (0.42)	6.91e-03 (2.4e-03)

Table 3: Evaluation of φ_i within **P2.** using credit card dataset with $n = 10$ data providers who each have a randomly sub-sampled dataset containing 100 training examples (Xu et al. 2021b).

et al. 2021b, Definition 3) (RVSV) and several real-world datasets for linear regression including used-car price prediction (Aditya 2019) and credit card fraud detection (Dal Pozzolo et al. 2014) where n data providers each owning a dataset (to estimate its RVSV). For **P3.** we consider (Sim et al. 2020, Equation 1) (CML) and hotel reviews sentiment prediction (Alam, Ryu, and Lee 2016) and Uber-lyft rides price prediction (BM 2018); in addition, we also consider (Wang et al. 2020, Definition 1) (FL) using two image recognition tasks (MNIST (LeCun et al. 1990) and CIFAR-10 (Krizhevsky, Sutskever, and Hinton 2012)) and two natural language processing tasks (movie reviews (Pang and Lee 2005) and Stanford Sentiment Treebank-5 (Kim 2014)). We partition the original dataset into n subsets, each owned by an agent i in FL/CML and we estimate each agent's contribution via the respective SV definitions. For **P4.** we follow (Lundberg and Lee 2017, Theorem 1) on several datasets including adult income (Kohavi and Becker 1996), iris (Fisher 1988), wine (Forina et al. 1991), and covertype (Blackard 1998) classification with different ML algorithms including kNN, logistic regression, SVM, and multi-layer perceptron (MLP). To ensure the experiments complete within reasonable time, we perform principal component analysis to obtain 7 principal components/features for computing $\boldsymbol{\phi}$.[10] For **hyperparameters**, since the largest n among these scenarios is 7, we set the budget $m = 1000$ and the bootstrapping of 300 evaluations (a total of 1300 evaluations for each baseline). We set $\xi =$ 1e-3 and vary $\alpha \in \{0, 2, 5, 100\}$ where 100 simulates $\alpha \to \infty$. Additional experimental details (datasets, ML models etc) are in (Zhou et al. 2023).

Evaluation and results. We examine the mean squared error (MSE) and mean absolute percentage error (MAPE) between $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$ for estimation accuracy, inversion counts N_{inv} and errors ϵ_{inv} for A3 and NL NSW (defined previously) for PDP. The inversion count $N_{\text{inv}} \coloneqq \sum_{i \neq j \in N} \mathbb{1}(\phi_i > \phi_j \cap \varphi_i < \varphi_j) + \mathbb{1}(\phi_i < \phi_j \cap \varphi_i > \varphi_j)$ is the number of inverted pairings in $\boldsymbol{\varphi}$ while $\epsilon_{\text{inv}} \coloneqq \sum_{i \neq j \in N} |\phi_i - \phi_j - (\varphi_i - \varphi_j)|$ is the sum of absolute errors (w.r.t. the true difference $\phi_i - \phi_j$). We present one set of average (and standard errors) over 5 repeated trials for **P2.** in Tab. 3 (the results for **P3.** and **P4.** and additional results are in (Zhou et al. 2023)). Overall, our method performs the best. While most methods perform competitively to ours w.r.t. MAPE, they are often worse (than ours) by an order of magnitude w.r.t. MSE. This is because our method explicitly addresses both the multiplicative and absolute errors (via $\xi = \epsilon_2/\epsilon_1$ in FS). Specifically, reducing absolute errors when $|\phi_i|$ is large (e.g., RVSV for **P2.** or (Sim et al. 2020, Equation 1) for **P3.** as both use the determinant of a large data matrix) is effective in reducing MSE. In our experiments, we find kernelSHAP underperforms others, which may be attributed to it having a larger (co-)variance,[11] empirically verified in (Covert and Lee 2021).

6 Discussion And Conclusion

We propose *probably approximate Shapley fairness* via a re-axiomatisation of Shapley fairness and subsequently exploit an error-aware *fidelity score* (FS) to provide a fairness guarantee with a polynomial (in n) budget complexity. We identify that jointly considering multiplicative and absolute errors (via their ratio ξ) is crucial in the quality of the fairness guarantee (which existing works did not do). Through analysing the effect of ξ on FS (used in our algorithm), we empirically find a suitable value for ξ. To achieve the fairness guarantee, we propose a novel *greedy active estimation* that integrates a greedy selection (which achieves a budget optimality) and active (permutation) selection via importance sampling. We identify that importance sampling can lead to poorer performance in practice as the necessary cardinality assumption may not be satisfied. To mitigate this, we describe a simple (via a single coefficient α) regularisation using a uniform Dirichlet prior, that interestingly unifies the frequentist and Bayesian approaches (its effectiveness is empirically verified). For future work, it is appealing to explore whether there exists a *biased* estimator with much lower variance to provide a similar/better fairness guarantee with a competitive budget complexity.

[10] We find if $n \geq 8$ features, the experiments take exceedingly long to complete due to the exponential complexity compounded further with the costly utility computation (Lundberg and Lee 2017, Equation 10).

[11] It is a co-variance matrix because kernelSHAP estimates the vector $\boldsymbol{\varphi}$ by solving a penalised regression.

Acknowledgements

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018). Xinyi Xu is also supported by the Institute for Infocomm Research of Agency for Science, Technology and Research (A*STAR).

References

Aditya. 2019. 100,000 UK used car dataset. URL https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes. Accessed: 2022-06-23.

Agarwal, A.; Dahleh, M.; and Sarkar, T. 2019. A marketplace for data: An algorithmic solution. In *Proceedings of the ACM Conference on Economics and Computation*, 701–726.

Agussurja, L.; Xu, X.; and Low, K. H. 2022. On the Convergence of the Shapley Value in Parametric Bayesian Learning Games. In *Proceedings of the 39th International Conference on Machine Learning*.

Alam, M. H.; Ryu, W.-J.; and Lee, S. 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339: 206–223.

Bahir, M.; Peleg, B.; Maschler, M.; and Peleg, B. 1966. A characterization, existence proof and dimension bounds for the kernel of a game. *Pacific Journal of Mathematics*, 18.

Bian, Y.; Rong, Y.; Xu, T.; Wu, J.; Krause, A.; and Huang, J. 2022. Energy-Based Learning for Cooperative Games, with Applications to Valuation Problems in Machine Learning. In *International Conference on Learning Representations*.

Blackard, J. 1998. Covertype. UCI Machine Learning Repository.

BM. 2018. Uber and Lyft dataset Boston, MA. URL https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma. Accessed: 2022-06-23.

Castro, J.; Gómez, D.; and Tejada, J. 2009. Polynomial Calculation of the Shapley Value Based on Sampling. *Computers & Operations Research*, 36(5): 1726–1730.

Castro, J.; Gómez, D.; Molina, E.; and Tejada, J. 2017. Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation. *Computers and Operations Research*, 82: 180–188.

Chalkiadakis, G.; Elkind, E.; and Wooldridge, M. 2011. *Computational Aspects of Cooperative Game Theory (Synthesis Lectures on Artificial Intelligence and Machine Learning)*. Morgan & Claypool Publishers, 1st edition.

Covert, I.; and Lee, S.-I. 2021. Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Dal Pozzolo, A.; Caelen, O.; Le Borgne, Y.-A.; Waterschoot, S.; and Bontempi, G. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10): 4915–4928.

Dalton, H. 1920. The Measurement of the Inequality of Incomes. *The Economic Journal*, 30(119): 348–361.

de Boer, J. F.; Cense, B.; Park, B. H.; Pierce, M. C.; Tearney, G. J.; and Bouma, B. E. 2003. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Optics Letters*.

de Clippel, G. 2010. Comment on "The Veil of Public Ignorance". Working paper.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least angle regression. *The Annals of Statistics*, 32(2).

Fisher, R. A. 1988. Iris. UCI Machine Learning Repository.

Forina, M.; et al. 1991. Wine. UCI Machine Learning Repository.

Ghorbani, A.; Kim, M. P.; and Zou, J. Y. 2020. A Distributional Framework for Data Valuation. In *Proceedings of the International Conference on Machine Learning*.

Ghorbani, A.; and Zou, J. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the International Conference on Machine Learning*, 2242–2251.

Ghorbani, A.; Zou, J.; and Esteva, A. 2021. Data Shapley Valuation for Efficient Batch Active Learning. arXiv:2104.08312.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Gurel, N. M.; Li, B.; Zhang, C.; Spanos, C. J.; and Song, D. 2019. Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms. In *Proceedings of the VLDB Endowment*, 1610–1623.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019. Towards Efficient Data Valuation Based on the Shapley Value. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 1167–1176.

Kaneko, M.; and Nakamura, K. 1979. The Nash Social Welfare Function. *Econometrica*, 47: 423–435.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kohavi, R.; and Becker, B. 1996. Adult. UCI Machine Learning Repository.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

Kwon, Y.; and Zou, J. 2022. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; and Jackel, L. 1990. Handwritten Digit Recognition with a Back-Propagation Network. In *Advances in Neural Information Processing Systems*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

Maleki, S.; Tran-Thanh, L.; Hines, G.; Rahwan, T.; and Rogers, A. 2013. Bounding the Estimation Error of Sampling-based Shapley Value Approximation With/Without Stratifying.

Mitchell, R.; Cooper, J.; Frank, E.; and Holmes, G. 2022. Sampling Permutations for Shapley Value Estimation. *Journal of Machine Learning Research*, 23(43): 1–46.

Nguyen, Q. P.; Low, K. H.; and Jaillet, P. 2022. Trade-off between Payoff and Model Rewards in Shapley-Fair Collaborative Machine Learning. In *Advances in Neural Information Processing Systems*.

Ohrimenko, O.; Tople, S.; and Tschiatschek, S. 2019. Collaborative Machine Learning Markets with Data-Replication-Robust Payments. In *Proceedings of the NeurIPS SGO & ML Workshop*.

Okhrati, R.; and Lipani, A. 2020. A Multilinear Sampling Algorithm to Estimate Shapley Values. In *Proceedings of the International Conference on Pattern Recognition*.

Owen, G. 1972. Multilinear Extensions of Games. *Source: Management Science*, 18: 64–79.

Pang, B.; and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedngs of the Annual Meeting of the Association for Computational Linguistics*.

Rose, A. 2012. *Vision: Human and Electronic*. Optical Physics and Engineering. Springer US. ISBN 9781468420395.

Rozemberczki, B.; Watson, L.; Bayer, P.; Yang, H.-T.; Kiss, O.; Nilsson, S.; and Sarkar, R. 2022. The Shapley Value in Machine Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Survey Track.

Sim, R. H. L.; Xu, X.; and Low, K. H. 2022. Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Survey Track.

Sim, R. H. L.; Zhang, Y.; Chan, M. C.; and Low, K. H. 2020. Collaborative Machine Learning with Incentive-Aware Model Rewards. In *Proceedings of the International Conference on Machine Learning*.

Song, T.; Tong, Y.; and Wei, S. 2019. Profit Allocation for Federated Learning. In *Proceedings of the IEEE International Conference on Big Data*, 2577–2586.

Street, N. 1995. UCI Machine Learning Repository Breast Cancer Wisconsin (Diagnostic) Data Set.

Sundararajan, M.; and Najmi, A. 2020. The Many Shapley Values for Model Explanation. In *Proceedings of the International Conference on Machine Learning*, 9269–9278. PMLR.

Tay, S.; Xu, X.; Foo, C. S.; and Low, K. H. 2022. Incentivizing Collaboration in Machine Learning via Synthetic Data Rewards. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9): 9448–9456.

Wang, T.; Rausch, J.; Zhang, C.; Jia, R.; and Song, D. 2020. A Principled Approach to Data Valuation for Federated Learning. *Lecture Notes in Computer Science*, 12500: 153–167.

Wu, Z.; Shu, Y.; and Low, K. H. 2022. DAVINZ: Data Valuation using Deep Neural Networks at Initialization. In *Proceedings of the International Conference on Machine Learning*.

Xu, X.; Lyu, L.; Ma, X.; Miao, C.; Foo, C. S.; and Low, K. H. 2021a. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Advances in Neural Information Processing Systems*.

Xu, X.; Wu, Z.; Foo, C. S.; and Low, K. H. 2021b. Validation free and replication robust volume-based data valuation. In *Advances in Neural Information Processing Systems*.

Zhou, Z.; Xu, X.; Sim, R. H. L.; Foo, C. S.; and Low, K. H. 2023. Probably Approximate Shapley Fairness with Applications in Machine Learning. arXiv:2212.00630.