Nonlinear Axiomatic Attribution for Cooperative Games

Weida Li[1] **Zhuanghua Liu**[1] **Yaoliang Yu**[2,3] **Bryan Kian Hsiang Low**[1]

[1] Department of Computer Science, National University of Singapore, Republic of Singapore
[2] School of Computer Science, University of Waterloo, Canada
[3] Vector Institute, Canada

Abstract

The Shapley value is a widely used concept in attribution problems, as it uniquely satisfies the axioms of linearity, consistency, equal treatment, and efficiency. Often, the inclusion AUC metric is used to evaluate the quality of player rankings, in order to identify positively participating players. However, it can be established that the Shapley value is not always reliable for this purpose. The core issue lies in its linearity: the Shapley value acts as a linear operator with an excessively large null space, which is likely to contain non-negligible perturbations that remain indistinguishable to the operator. To address this limitation, we explore the design of nonlinear axiomatic attribution methods. Inspired by the least core, which is a popular nonlinear substitute for the Shapley value, we introduce a class of nonlinear attribution methods that retain the remaining necessary axioms. Each method yields a contribution vector that is the unique optimal solution to a minimization problem, which aims to approximate utility functions as faithfully as possible. In terms of the inclusion AUC metric, our experiments demonstrate the potential effectiveness of these methods compared to Shapley value variants that relax only the efficiency axiom. Our code is available at https://github.com/watml/nonlinear-axiom.

1 INTRODUCTION

Originally, the concept of the Shapley value was introduced by Shapley [1953] to define a fair allocation of contributions among n players who participate in a cooperative game, represented by a utility function $U\colon 2^{[n]} \to \mathbb{R}$, where $[n] := \{1, 2, \dots, n\}$. It is considered fair because it uniquely satisfies the axioms of linearity, consistency, equal treatment, and efficiency. The effectiveness of the Shapley value and its variants obtained by relaxing the efficiency axiom has been demonstrated by Jia et al. [2019b], Kwon and Zou [2022a], Li and Yu [2023], Wang and Jia [2023] in data attribution, where data are treated as players and $U(S)$ is defined as the performance of a model trained on the subset S of available training data. One such variant, namely the Banzhaf value [Banzhaf III, 1965], was used in context attribution by Cohen-Wang et al. [2024], where $U(S)$ represents the probability of a statement generated by a large language model when only the parts of the context text prescribed by S are used. Its popularity has also been prominently observed in feature attribution, where $U(S)$ is the predicted value of the model when features outside the subset S are considered missing [e.g., Lundberg and Lee, 2017, Kwon and Zou, 2022b].

However, Kumar et al. [2020], Yan and Procaccia [2021] have raised concerns about the imposed linearity axiom, as its necessity remains unclear. Meanwhile, Bilodeau et al. [2024], Wang et al. [2024] have theoretically established that the use of the Shapley value can be unreliable by exploiting linearity. Accordingly, a nonlinear alternative, the least core [Shapley, 1971], has attracted attention [Yan and Procaccia, 2021, Gemp et al., 2024]. In general, the least core may contain infinitely many contribution allocations, so we instead consider the egalitarian least core [Arin et al., 2008, Yan and Procaccia, 2021].[1] Still, the egalitarian least core is constrained by the efficiency axiom, which has been empirically shown to be unnecessary, and even unfavorable, by Kwon and Zou [2022a,b]. By relaxing only the efficiency axiom, one arrives at the family of semi-values [Dubey et al., 1981], which includes Beta Shapley values [Kwon and Zou, 2022a] and weighted Banzhaf values [Li and Yu, 2023], uniquely characterized by the axioms of linearity, consistency, equal treatment, and monotonicity.[2] In this work, we

[1] The least core is a convex subset of $\mathbb{R}^n$, and the egalitarian least core is defined as the element in the least core with the smallest ℓ_2 norm.

[2] The Shapley value also satisfies the monotonicity axiom.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3581–3603.

theoretically explore the design of attribution methods by relaxing only the axioms of linearity and efficiency.

Our contributions. Often, the quality of player rankings is evaluated using the inclusion Area Under the Curve (AUC) metric, when the goal is to identify positively contributing players [Petsiuk et al., 2018, Kwon and Zou, 2022a, Covert et al., 2023]. Under this metric, it can be demonstrated (Lemmas 3.2 and 3.3) that applying the egalitarian least core to rank players is equivalent to approximately maximizing the inclusion AUC by substituting an additive approximation of the utility function U. This perspective follows from the imposed efficiency axiom. Likewise, the use of semi-values in player ranking can be interpreted in a similar way (Corollary 3.4).

- We then present a theoretical result demonstrating that the Shapley value is not reliable for maximizing the inclusion AUC, due to a vulnerability introduced by the linearity axiom (Theorem 4.1).
- In Section 4.2, based on the Shapley value, we show via a pathological example that using the sum of contribution vectors of different utility functions as the contribution vector of their sum can lead to undesirable behavior w.r.t. maximizing the inclusion AUC.

These observations motivate us to explore nonlinear attribution methods that do not impose the linearity axiom, yet still satisfy the remaining desired axioms. In particular, the interpretation of applying the egalitarian least core as well as semi-values to rank players points us toward starting from a faithful additive approximation of utility functions. Consequently,

- We introduce a mathematically grounded class of attribution methods that relaxes only the linearity and efficiency axioms (Theorem 4.3);[3]
- We also lay out a theoretical ground for approximately computing the egalitarian least core and our introduced nonlinear attribution method ϕ^{∞} (Theorems 4.5 and 4.6);

Finally, we conduct experiments to verify the practical effectiveness of the proposed methods in improving player ranking quality under the inclusion AUC metric.

2 BACKGROUND

2.1 NOTATION

Let $[n] \coloneqq \{1, 2, \ldots, n\}$ denote a set of players, where $n \geq 2$. For each n, $\mathcal{G}_n \coloneqq \{U_n \colon 2^{[n]} \to \mathbb{R}\}$ refers to the set of utility functions over n players. Throughout the paper, the subscript n indicates the number of players. We then define the space of all such games as $\mathcal{G} \coloneqq \bigcup_{n \geq 2} \mathcal{G}_n$. An attribution method ϕ maps each utility function $U_n \in \mathcal{G}$ to a contribution vector $\phi(U_n) \in \mathbb{R}^n$, where $\phi_i(U_n)$ measures the contribution of the i-th player in the cooperative game defined by U_n. The all-one vector in $\mathbb{R}^m$ is denoted by $\mathbf{1}_m$. Given $\mathbf{x} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, we define an additive utility function $A_n(\cdot; \mathbf{x}, b) \in \mathcal{G}_n$ by letting

$$A_n(S; \mathbf{x}, b) \coloneqq b + \sum_{i \in S} x_i \quad \text{for every} \quad S \subseteq [n],$$

where by default an empty sum is taken as 0. Plus, $[\![\cdot]\!]$ denotes the indicator function.

2.2 AXIOMS IN COOPERATIVE GAME THEORY

In what follows, we present a list of axioms commonly used in the literature [e.g., Shapley, 1953, Dubey et al., 1981, Weber, 1988, Li and Yu, 2023].

- **Linearity**: $\phi(c \cdot U_n + U'_n) = c \cdot \phi(U_n) + \phi(U'_n)$ for every $U_n, U'_n \in \mathcal{G}$ and $c \in \mathbb{R}$.
- **Consistency**: Suppose there exists $c \in \mathbb{R}$ such that $U_{n+1}(S) = U_n(S \cap [n]) + c \cdot [\![n+1 \in S]\!]$ for every $S \subseteq [n+1]$. Then, (i) $\phi_{n+1}(U_{n+1}) = c$, and (ii) $\phi_i(U_n) = \phi_i(U_{n+1})$ for every $i \in [n]$.
- **Equal Treatment**: For some $i, j \in [n]$, if $U_n(S \cup \{i\}) = U_n(S \cup \{j\})$ for every $S \subseteq [n] \setminus \{i, j\}$, then $\phi_i(U_n) = \phi_j(U_n)$.
- **Monotonicity**: For some $i \in [n]$, (i) if $U_n(S \cup \{i\}) \geq U_n(S)$ for every $S \subseteq [n] \setminus \{i\}$, then $\phi_i(U_n) \geq 0$; (ii) if $U_n(S \cup \{i\}) \leq U_n(S)$ for every $S \subseteq [n] \setminus \{i\}$, then $\phi_i(U_n) \leq 0$.
- **Efficiency**: $\sum_{i \in [n]} \phi_i(U_n) = U_n([n]) - U_n(\emptyset)$ for every $U_n \in \mathcal{G}$.
- **Translation Invariance**: $\phi(U_n + C_n) = \phi(U_n)$ whenever $C_n \in \mathcal{G}$ is a constant utility function.

We note that the consistency axiom implies both the dummy axiom used by Weber [1988] and the projection axiom introduced by Dubey et al. [1981]. Besides, translation invariance is often implied by linearity and other axioms. In other words, translation invariance may be viewed as a weak version of linearity.

2.3 SEMI-VALUES

The family of semi-values is uniquely characterized by all the above axioms except efficiency [Dubey et al., 1981]. Specifically, each semi-value corresponds to a Borel probability measure μ over the interval $[0, 1]$. Accordingly, we

[3] As an unexpected finding, we also concurrently developed another class of such nonlinear axiomatic attribution methods via optimizing the inclusion AUC metric [Li et al., 2026].

use ϕ^{μ} to denote the semi-value parameterized by μ. The contribution of the i-th player in a game U_n is given by

$$\phi_i^{\mu}(U_n) = \sum_{S \subseteq [n] \setminus i} p_{n,|S|+1}^{\mu} [U_n(S \cup \{i\}) - U_n(S)]$$

$$\text{where } p_{n,k}^{\mu} = \int_0^1 t^{k-1}(1-t)^{n-k} \mathrm{d}\mu(t).$$

When the efficiency axiom is additionally imposed, μ is uniquely determined as the uniform distribution on $[0,1]$, resulting in the well-known Shapley value. For the weighted Banzhaf value $\mathrm{WB}(\tau)$, where $\tau \in (0,1)$, the corresponding measure μ is the Dirac delta distribution δ_τ. For the Beta Shapley value $\mathrm{Beta}(\alpha, \beta)$, where $\alpha, \beta \geq 1$, the measure μ is defined by

$$\mu(S) \propto \int_S t^{1-\beta}(1-t)^{1-\alpha} \mathrm{d}t$$

for every measurable subset S of $[0,1]$. In particular, $\mathrm{Beta}(1,1)$ corresponds to the Shapley value. Empirically, Kwon and Zou [2022a,b], Li and Yu [2023] demonstrate that Beta Shapley values and weighted Banzhaf values often perform better in downstream tasks, challenging the necessity of the efficiency axiom. Heuristically, when only player rankings are of interest, the efficiency axiom, as a normalization step, is often regarded as redundant.

3 MOTIVATIONS

3.1 THE EGALITARIAN LEAST CORE

Following Yan and Procaccia [2021], the least core $\mathrm{LC}(U_n)$ of a utility function U_n is a subset of $\mathbb{R}^n$ containing all the optimal solutions of $\mathbf{x}$ to the minimization problem

$$\begin{aligned} &\underset{\mathbf{x} \in \mathbb{R}^n, e \in \mathbb{R}}{\text{minimize}} \ e \\ &\text{s.t. } U_n(S) - A_n(S; \mathbf{x}, U_n(\emptyset)) \leq e, \ \forall\, S \subsetneq [n] \\ &\quad \text{and } A_n([n]; \mathbf{x}, U_n(\emptyset)) = U_n([n]). \end{aligned} \tag{1}$$

According to Blum and Oettli [1972], it is clear that the least core $\mathrm{LC}(U_n)$ is always non-empty. The egalitarian least core ϕ^{ELC} is then defined as the least-norm element in $\mathrm{LC}(U_n)$. Precisely,

$$\phi^{\mathrm{ELC}}(U_n) \coloneqq \underset{\mathbf{x} \in \mathrm{LC}(U_n)}{\arg\min} \|\mathbf{x}\|_2.$$

The least norm solution is critical for ensuring that ϕ^{ELC} satisfies the axiom of equal treatment [Yan and Procaccia, 2021]. To the best of our knowledge, no existing work has established whether ϕ^{ELC} satisfies the other axioms listed above. Therefore, for completeness, we analyze the axiomatic properties of the egalitarian least core.

Theorem 3.1. *The egalitarian least core ϕ^{ELC} is nonlinear and satisfies the axioms of consistency, equal treatment, monotonicity, efficiency, and translation invariance.*

This result indicates that the egalitarian least core relaxes only the linearity axiom. Interestingly, we observe that the efficiency axiom effectively transforms the one-sided error bounds in the problem (1) into two-sided ones. The following lemma makes this precise.

Lemma 3.2. *Given a utility function $U_n \in \mathcal{G}$, a vector $\mathbf{x}' \in \mathrm{LC}(U_n)$ if and only if $\mathbf{x}'$ is an optimal solution to the problem*

$$\begin{aligned} &\underset{\mathbf{x} \in \mathbb{R}^n, e \in \mathbb{R}}{\text{minimize}} \ e \\ &\textit{s.t.} \ -e \leq A_n(S; \mathbf{x}, U_n(\emptyset)) - U_n(S), \ \forall\, S \subsetneq [n] \\ &A_n(S; \mathbf{x}, U_n(\emptyset)) - U_n(S) \leq e + B(S), \ \forall\, S \subsetneq [n] \\ &\quad \textit{and } \sum_{i \in [n]} x_i = U_n([n]) - U_n(\emptyset) \end{aligned} \tag{2}$$

where $B(S) \coloneqq U_n([n]) + U_n(\emptyset) - U_n(S) - U_n([n] \setminus S)$.

Proof. Suffice it to demonstrate that $A_n(S; \mathbf{x}, U_n(\emptyset)) \leq e + B(S)$ is equivalent to

$$U_n([n] \setminus S) - A_n([n] \setminus S; \mathbf{x}, U_n(\emptyset)) \leq e.$$

Using the efficiency constraint,

$$\begin{aligned} &A_n([n] \setminus S; \mathbf{x}, U_n(\emptyset)) + A_n(S; \mathbf{x}, U_n(\emptyset)) \\ &\quad = U_n([n]) + U_n(\emptyset), \end{aligned}$$

which leads to

$$\begin{aligned} &U_n([n] \setminus S) - A_n([n] \setminus S; \mathbf{x}, U_n(\emptyset)) \\ &= A_n(S; \mathbf{x}, U_n(\emptyset)) - U_n(S) - B(S). \end{aligned}$$

□

In summary, the egalitarian least core produces an additive utility function $A_n(\cdot; \phi^{\mathrm{ELC}}(U_n), U_n(\emptyset))$ that seeks to approximate U_n under the problem (2).

3.2 EVALUATING PLAYER RANKINGS

In data attribution [Wang et al., 2024], the utility function $U_n(S)$ is typically defined as the performance of a model trained on the subset S of available training data. Then, the grand set of players $[n]$ corresponds to the set of training data, and $\phi_i(U_n)$ quantifies the contribution of the i-th data point to the performance of trained models. In this context, the goal is to identify positively contributing data points, which is closely related to maximizing the inclusion AUC, defined as

$$\begin{aligned} &\underset{\pi \in \Pi_n}{\text{maximize}} \ \mathrm{AUC}(\pi; U_n) \coloneqq \sum_{k=1}^{n} U_n(S_k(\pi)) \\ &\text{where } S_k(\pi) \coloneqq \{\pi(1), \pi(2), \ldots, \pi(k)\}. \end{aligned} \tag{3}$$

Here, Π_n denotes the set of all permutations of $[n]$. We note that maximizing the inclusion AUC also applies to context attribution [Cohen-Wang et al., 2024], where the aim is to identify the influential parts of the context that generate a statement. In general, solving the problem (3) is computationally challenging due to the factorial number of possible permutations. Nevertheless, the problem becomes tractable when the utility function is additive.

Lemma 3.3. *Let $\mathbf{x} \in \mathbb{R}^n$, and define $\pi_{\mathbf{x}}$ as the permutation satisfying $x_{\pi_{\mathbf{x}}(j)} \geq x_{\pi_{\mathbf{x}}(j+1)}$ for every $1 \leq j < n$. Then, $\pi_{\mathbf{x}}$ is optimal to the problem*

$$\underset{\pi \in \Pi_n}{\text{maximize}} \ \text{AUC}(\pi; A_n(\cdot; \mathbf{x}, b))$$

where $b \in \mathbb{R}$ is arbitrary.

Proof. This result follows immediately from the observation

$$\text{AUC}(\pi; A_n(\cdot; \mathbf{x}, b)) = \sum_{k=1}^{n} (n+1-k) \cdot x_{\pi(k)}$$

and the rearrangement inequality. □

Combining Lemmas 3.2 and 3.3, we interpret the mechanism of the egalitarian least core in feature ranking as follows:

- Obtain the best additive approximation of U_n by solving the minimization problem (2), yielding the additive utility function $A_n(\cdot; \phi^{\text{ELC}}(U_n), U_n(\emptyset))$.
- Substitute the obtained $A_n(\cdot; \phi^{\text{ELC}}(U_n), U_n(\emptyset))$ into the maximization problem (3) to derive an approximate optimal ranking.

Meanwhile, semi-values admit a similar interpretation. Specifically, Li and Yu [2024, Theorem 2] proved that there exists a function $T^\mu : \mathcal{G} \to \mathbb{R}$ such that

$$\phi^\mu(U_n) = \mathbf{x}^*_\mu(U_n) + T^\mu(U_n) \cdot \mathbf{1}_n$$

where $(\mathbf{x}^*_\mu(U_n), b^*_\mu(U_n))$ is the unique optimal solution to the least square problem

$$\underset{\mathbf{x} \in \mathbb{R}^n, b \in R}{\text{minimize}} \sum_{\emptyset \subsetneq S \subsetneq [n]} p^\mu_{n-1,|S|} [U_n(S) - A_n(S; \mathbf{x}, b)]^2.$$

This immediately leads to the following corollary.

Corollary 3.4. *Given a utility function $U_n \in \mathcal{G}$, let π_μ be a ranking of $[n]$ such that $\phi^\mu_{\pi_\mu(j)}(U_n) \geq \phi^\mu_{\pi_\mu(j+1)}(U_n)$ for every $1 \leq j < n$. Then, π_μ is an optimal solution to*

$$\underset{\pi \in \Pi_n}{\text{maximize}} \ \text{AUC}[\pi; A_n(\cdot; \mathbf{x}^*_\mu(U_n), b^*_\mu(U_n))].$$

This corollary is immediate by observing that the induced player rankings of $\phi^\mu(U_n)$ and $\mathbf{x}^*_\mu$ are the same.

In summary, many popular axiomatic attribution methods for ranking players (e.g., Beta Shapley values, weighted Banzhaf values, and the egalitarian least core) can be viewed as implicitly learning an additive approximation to the true utility function U_n, which is then used to derive an approximate optimal ranking. This motivates us to next explore nonlinear, non-efficient attribution methods through the same lens of approximation.

4 MAIN RESULTS

4.1 HOW UNRELIABLE IS THE SHAPLEY VALUE

Before proceeding, we demonstrate that the Shapley value may fail to distinguish between utility functions in terms of approximately maximizing the inclusion AUC.

Theorem 4.1. *Let Π_n be the set of all permutations of $[n]$ with $n \geq 3$, and let ϕ^{Shap} denote the Shapley value. Then, Π_n can be partitioned into $\{\Pi_{n,j}\}_{j \in \mathcal{J}}$ with $|\mathcal{J}| \geq 2$ such that, for every $\Pi_{n,j}$ and every $U_n \in \mathcal{G}_n$, there exists a perturbation $U'_n \in \mathcal{G}_n$ satisfying*

$$\phi^{\text{Shap}}(U_n) = \phi^{\text{Shap}}(U_n + U'_n),$$
$$\textit{yet} \ \ \arg\max_{\pi \in \Pi_n} \text{AUC}(\pi; U_n + U'_n) \subseteq \Pi_{n,j}.$$

Theorem 4.1 states that, preserving the same Shapley value, the AUC-optimal permutation can be manipulated to lie in any $\Pi_{n,j}$. Specifically, the partition in Theorem 4.1 is constructed by exploiting the null space of the Shapley value. In other words, the linearity axiom induces an excessively large null space that can be detrimental. We note that this is not the only negative perspective established so far. For example, Wang et al. [2024, Theorem 2] showed that the Shapley value is not reliable for maximizing the utility $U(S)$. On the other hand, Bilodeau et al. [2024, Theorem B.3] demonstrated that the Shapley value fails to distinguish local behaviors of different functions. In particular, all these negative results rely on exploiting the null space induced by the linearity axiom.

4.2 A PATHOLOGICAL EXAMPLE INDUCED BY LINEARITY

The inarguable merit of linearity is that it divides the calculation of $\phi(U + U')$ into $\phi(U) + \phi(U')$. However, we demonstrate that such a convenience comes with pathological examples in maximizing the problem (3). Let $2^{[4]}$ be ordered as

$$\emptyset, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \{4\}, \{1,4\},$$
$$\{2,4\}, \{1,2,4\}, \{3,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\},$$

which is the binary ordering used in Grabisch et al. [2000]. Then, every utility function in $\mathcal{G}_4$ can be expressed as a vector. Let

$$\begin{aligned} U_4 &= (4, -3, -4, -8, 2, 6, 3, -8, \\ &\quad\ 4, -10, -3, 8, 5, 1, 4, -3), \\ U_4' &= (-5, -2, -7, 7, -10, 4, -4, 9, \\ &\quad\ -1, -6, -1, 1, -3, -2, -10, -5). \end{aligned}$$

Particularly, they are constructed such that $\pi^{\mathrm{Shap}}(U_4)$ and $\pi^{\mathrm{Shap}}(U_4')$ each exactly maximize the problem (3), where π^{Shap} denotes the player ranking induced by the Shapley value. However, one can verify that $\pi^{\mathrm{Shap}}(U_4 + U_4')$ instead *minimizes* the inclusion AUC. This pathological example clearly demonstrates the unexpected behavior that arises from simply adding the Shapley values of U_4 and U_4' when maximizing the inclusion AUC. Moreover, there exist many such examples, as the one presented here was instantly found via random search.

The above example raises a concern about the common practice of using the sum of contribution vectors of different utility functions as the contribution vector of their sum. This phenomenon also serves as our motivation to explore what nonlinear axiomatic attribution methods would look like.

4.3 NONLINEAR ATTRIBUTION METHODS

Looking at the problem (2) of the least core, the term B appears to hinder faithful approximation of U_n. Recall that this term arises from enforcing the efficiency axiom. To relax this axiom, it is natural to allow two-sided approximation errors for each subset S, i.e., bounding $A_n(S; \mathbf{x}, b) - U_n(S)$ from both sides. Moreover, to better fit U_n, we treat the bias term b as a variable to be optimized, rather than fixing it straightforwardly to $U_n(\emptyset)$. All in all, we consider the minimization problem below that attempts to faithfully approximate the utility function using an additive one:

$$\underset{\mathbf{x}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \ \|\mathbf{A}\mathbf{x} + b \cdot \mathbf{1}_{2^n} - \mathbf{V}(U_n)\|_p \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{2^n \times n}$, and $\mathbf{V}(U_n) \in \mathbb{R}^{2^n}$ contains $U_n(S)$ for all subsets $S \subseteq [n]$. Each row A_S of $\mathbf{A}$ is the indicator vector of the subset S, i.e., $A_{S,i} = 1$ if $i \in S$, and 0 otherwise. Note that when $p = \infty$, the problem (4) coincides with the problem (2) with the efficiency axiom removed and $B(S) \equiv 0$. We note that Charnes et al. [1988] considered a similar approximation with $p \in \{1, 2, \infty\}$.

Let $\mathrm{OS}_p(U_n)$ denote the set of all optimal solutions of $\mathbf{x}$ to the minimization problem (4). Then, We define the p-norm attribution method $\phi^p(U_n)$ by selecting the minimum ℓ_2-norm solution among $\mathrm{OS}_p(U_n)$:

$$\phi^p(U_n) \coloneqq \underset{\mathbf{x}\in\mathrm{OS}_p(U_n)}{\arg\min} \ \|\mathbf{x}\|_2.$$

Algorithm 1 Approximation of $\phi^p(U_n)$

Input: Utility function U_n, regularization parameter $\eta = 10^{-12}$, $p \in (1, \infty) \cup \{\infty, \mathrm{ELS}\}$, and sample budget B

Output: Approximation $\hat{\mathbf{x}}^* \in \mathbb{R}^n$ of $\phi^p(U_n)$

$\mathcal{S} \leftarrow [\,]$
for $b = 1, 2, \ldots, B$ **do**
  Uniformly sample a subset $S_b \subseteq [n]$ not already in $\mathcal{S}$;
  if $p = \mathrm{ELS}$ **then**
    Resample S_b if $S_b \in \{\emptyset, [n]\}$
  $\mathcal{S}.\mathrm{append}(S_b)$
if $p \in (1, \infty) \cup \{\infty\}$ **then**
  Construct $\mathbf{A} \in \mathbb{R}^{B \times n}$ and $\mathbf{V} \in \mathbb{R}^B$ using $\mathcal{S}$ and U_n
  $\eta \leftarrow 0$ if $p \in (1, \infty)$
  $\hat{\mathbf{x}}^* \leftarrow \mathrm{minimize}_{\mathbf{x}\in\mathbb{R}^n, b\in\mathbb{R}} \ \|\mathbf{A}\mathbf{x} + b \cdot \mathbf{1}_B - \mathbf{V}\|_p + \eta \cdot \|\mathbf{x}\|_2^2$
else if $p = \mathrm{ELC}$ **then**
  $\hat{\mathbf{x}}^* \leftarrow \mathrm{minimize}_{\mathbf{x}\in\mathbb{R}^n, e\in\mathbb{R}} \ e + \eta \cdot \|\mathbf{x}\|_2^2$ subject to $U_n(S) - U_n(\emptyset) - \sum_{i\in S} x_i \le e$ for every $S \in \mathcal{S}$, $0 \le e$, and $\sum_{i\in[n]} x_i = U_n([n]) - U_n(\emptyset)$
return $\hat{\mathbf{x}}^*$

For $p \in (1, \infty)$, the problem admits a unique solution since the objective is strictly convex. However, for $p \in \{1, \infty\}$, multiple optimal solutions may exist. As an example for $p = \infty$, consider $n = 3$ and define U_3 such that

$$U(S) \coloneqq \begin{cases} 0, & S = \emptyset, \\ 0.5, & S \in \{\{2\}, \{3\}\}, \\ 1, & \text{otherwise.} \end{cases}$$

Then, one can verify that the solution of $b = 0.25$ and $\mathbf{x} = (0.5, 0.25 + t, 0.25 - t)$ is optimal for every $t \in [-0.25, 0.25]$.

Proposition 4.2. *The 1-norm attribution method ϕ^1 does not satisfy the consistency axiom.*

Such a failure of the 1-norm attribution method ϕ^1 arises because $\mathrm{minimize}_{b\in\mathbb{R}} \sum_{j=1}^m |b - y_j|$ does not necessarily yield a unique solution. We provide a concrete counterexample to prove Proposition 4.2.

Proof. Define $U_2 \in \mathcal{G}_2$ by letting

$$U_2(\emptyset) = 0, \ U_2(\{1\}) = 3, \ U_2(\{2\}) = 2, \ U_2(\{1, 2\}) = 1.$$

Then, $\phi^1(U_2) = (0, 0)$ as the solution of $b = 1.5$ and $\mathbf{x} = (0, 0)$ is optimal to the problem

$$\underset{\mathbf{x}\in\mathbb{R}^2, b\in\mathbb{R}}{\text{minimize}} \sum_{S\subseteq[2]} |U_2(S) - b - \sum_{i\in S} x_i|.$$

The optimality is due to that the problem is convex and $\mathbf{0}$ is a subgradient at that point. In particular, $(b, 0, 0)$ with

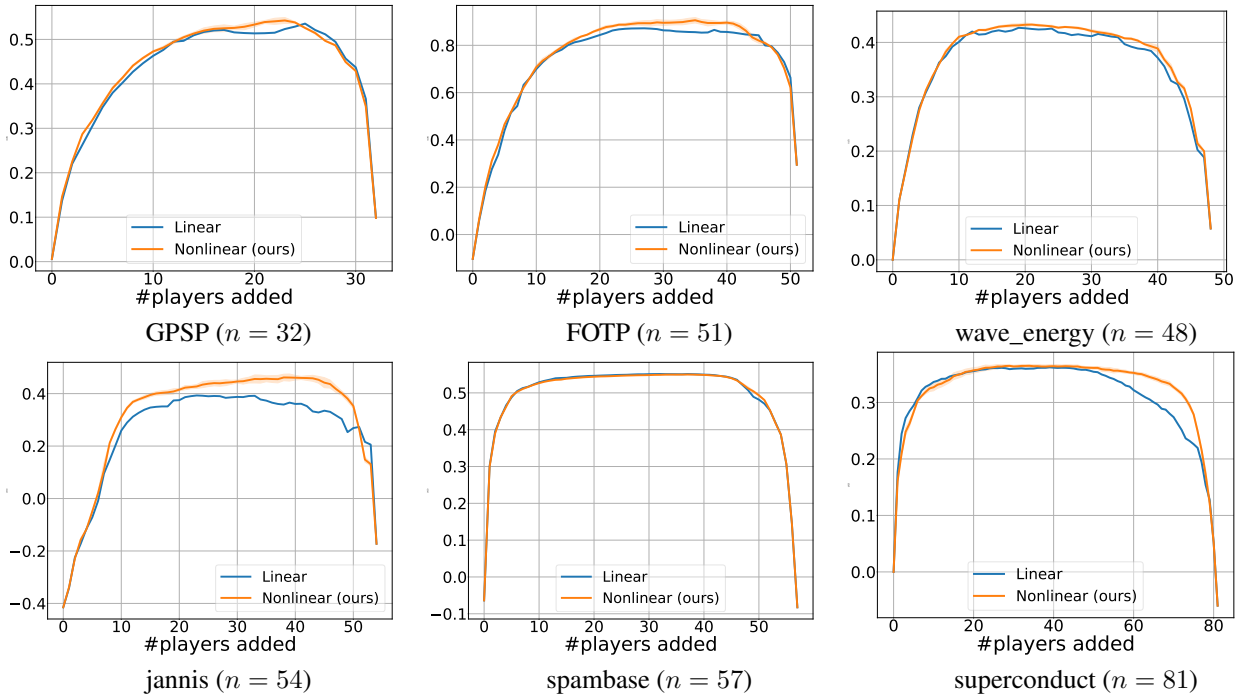

Figure 1: Comparison of attribution methods on six datasets for player ranking. A larger area under the curve indicates better performance. Since all utility functions contain more than 30 players, nonlinear attribution methods are approximated by sampling $\#\text{players} \times 1,000$ subsets, with mean and standard deviation reported over five random seeds. Linear methods are computed exactly. The linear and nonlinear curves shown are the best among 11 and 6 candidates respectively under the inclusion AUC metric.

$b \in [1, 2]$ is optimal to the problem. Then, we construct a utility function $U_3 \in \mathcal{G}_3$ by letting

$$U_3(S) := \begin{cases} U_2(S), & 3 \notin S, \\ U_2(S) + 1, & \text{otherwise.} \end{cases}$$

Note that

$$\min_{\mathbf{x} \in \mathbb{R}^3, b \in \mathbb{R}} \sum_{S \subseteq [3]} |U_3(S) - b - \sum_{i \in S} x_i| = 2e.$$

Then, one can verify that the solution of $b = 2$ and $\mathbf{x} = (0, 0, 0)$ is optimal, which suggests that $\phi^1(U_3) = (0, 0, 0)$. With this example, it is clear that ϕ^1 violates the axiom of consistency. □

Nevertheless, for the remaining values of p, the resulting attribution methods are axiomatically sound.

Theorem 4.3. *For $p \in (1, \infty) \cup \{\infty\}$, the p-norm attribution method ϕ^p satisfies the axioms of consistency, equal treatment, monotonicity and translation invariance. When $n > 2$, ϕ^p is linear if and only if $p = 2$.*

Possible Weighted Extensions. For $p \in (1, \infty)$, we can define weighted versions of the p-norm objective while retaining the axiomatic guarantees of Theorem 4.3. Given $\omega \in (0, 1)$ and $p \in (1, \infty)$, define $\phi^{p,\omega}(U_n)$ as the unique optimal solution of $\mathbf{x}$ to the problem

$$\underset{\mathbf{x} \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \sum_{S \subseteq [n]} \omega^{|S|}(1 - \omega)^{n - |S|} \left| A_n(S; \mathbf{x}, b) - U_n(S) \right|^p .$$

Then, $\phi^{2,\omega}$ is the weighted Banzhaf value parameterized by ω [Marichal and Mathonet, 2011, Li and Yu, 2023]. Though we did not find a rigorous way to prove that $\phi^{p,\omega}$ is linear if and only if $p = 2$, we conjecture that this claim holds true.

Corollary 4.4. *For every $\omega \in (0, 1)$ and every $p \in (1, \infty)$, the weighted p-norm attribution method $\phi^{p,\omega}$ satisfies the axioms of consistency, equal treatment, monotonicity and translation invariance.*

4.4 OPTIMIZATION

Note that when n is sufficiently large, all of the mentioned attribution methods can only be approximated, unless the utility function U_n exhibits certain favorable properties [e.g., Jia et al., 2019a, Lundberg et al., 2020].

For $p \in (1, \infty)$, $\phi^p(U_n)$ can be effectively approximated by solving the minimization problem (4), since its solution is unique. However, this is not the case for ϕ^∞, as it is defined as the optimal solution with the smallest ℓ_2-norm among potentially many minimizers. To facilitate the approximation of ϕ^∞, we establish the following theoretical result.

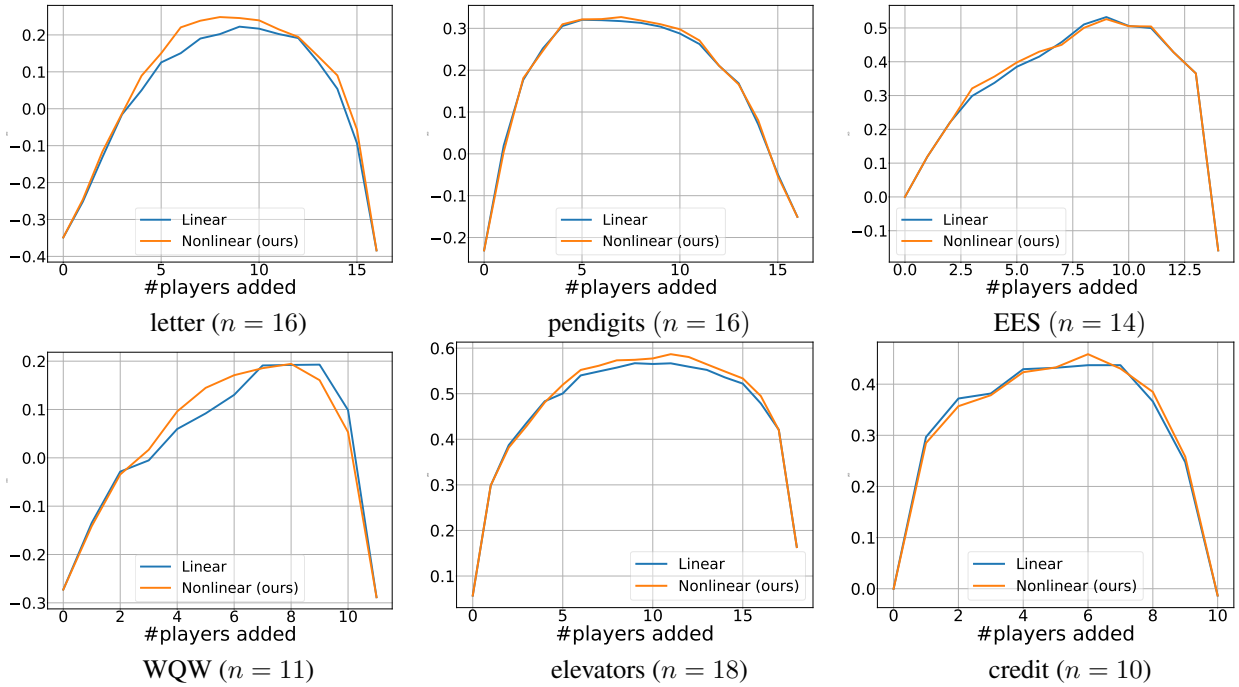

Figure 2: Comparison of attribution methods on six datasets for feature ranking. A larger area under the curve indicates better performance. Since all utility functions contain fewer than 20 players, all subsets are used for nonlinear methods and exact computations are performed for linear methods. The linear and nonlinear curves shown are the best among 11 and 6 candidates respectively under the inclusion AUC metric.

Theorem 4.5. *For every $\eta > 0$, let $\phi^{\infty}(U_n; \eta)$ denote an optimal solution of $\mathbf{x}$ to*

$$\underset{\mathbf{x}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \ \|\mathbf{A}\mathbf{x} + b \cdot \mathbf{1}_{2^n} - \mathbf{V}(U_n)\|_{\infty} + \eta \cdot \|\mathbf{x}\|_2^2.$$

Then, we have $\phi^{\infty}(U_n; \eta) \to \phi^{\infty}(U_n)$ as $\eta \to 0$.

The key to proving this result is that the constraint $\|\mathbf{A}\mathbf{x} - b \cdot \mathbf{1}_n - \mathbf{V}(U_n)\|_{\infty} \leq e$ implies boundedness of $(\mathbf{x}, b)$. While previous studies have focused on approximating the least core, it remains unclear how to approximate the egalitarian least core ϕ^{ELC}. Nevertheless, the regularization technique used for approximating ϕ^{∞} can be extended to ϕ^{ELC} as well.

Theorem 4.6. *For every $\eta > 0$, let $\phi^{\mathrm{ELC}}(U_n; \eta)$ denote the unique optimal solution of $\mathbf{x}$ to*

$$\underset{\mathbf{x}\in\mathbb{R}^n, e\in\mathbb{R}}{\text{minimize}} \ e + \eta \cdot \|\mathbf{x}\|_2^2$$

$$\text{subject to} \ \ U_n(S) - A_n(S; \mathbf{x}, U_n(\emptyset)) \leq e \ \text{ for every } S \subsetneq [n]$$
$$\text{and} \ \ A_n([n]; \mathbf{x}, U_n(\emptyset)) = U_n([n]).$$

Then, we have $\phi^{\mathrm{ELC}}(U_n; \eta) \to \phi^{\mathrm{ELC}}(U_n)$ as $\eta \to 0$.

Accordingly, we propose Algorithm 1 to approximate the attribution methods of interest when exact computation is infeasible. Specifically, all the minimization problems we solve are convex, and thus off-the-shelf convex solvers can be directly applied.

5 EXPERIMENTS

The goal of our experiments is to evaluate whether the nonlinear attribution methods introduced in this work outperform semi-values in ranking players. According to Theorem 4.3, the considered nonlinear attribution methods include ϕ^p with $p \in \{1.5, 2.5, 5, 10, \infty\}$ as well as ϕ^{ELC}. We use the CVXPY library with the CLARABEL solver to solve the minimization problems in Algorithm 1 [Diamond and Boyd, 2016, Agrawal et al., 2018].

Utility functions. The utility function we employ is $U_f^{\mathbf{x}}(S) := \mathbb{E}_{\mathbf{X}_{[n]\setminus S}}[f(\mathbf{x}_S, \mathbf{X}_{[n]\setminus S})]$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a trained gradient boosted tree (GBT) with five estimators, $\mathbf{x} \in \mathbb{R}^n$ is an input instance, and $\mathbf{x}_S$ is the restriction of $\mathbf{x}$ on S. Then, the number of players corresponds to the number of features. We follow the path-dependent definition proposed by Lundberg et al. [2020, Algorithm 1]. For this type of utility functions, the corresponding semi-values can be computed in polynomial time [Karczmarz et al., 2022, Yu et al., 2022, Muschalik et al., 2024], and we use the numerically stable version developed by Li et al. [2026]. In general, $\phi^{\mu}(f_{\mathbf{x}})$ can only be approximated, and our rationale is to avoid such approximation error. All GBTs are trained using the scikit-learn library [Pedregosa et al., 2011]. We train one GBT for each dataset. To account for the variability of $U_f^{\mathbf{x}}$, for each trained GBT f, the reported curve is averaged

Table 1: Average inclusion AUC scores of various attribution methods. The best results are shown in bold, and the second-best results are underlined. Note that Beta Shapley reports the best one over 10 candidates, as described in (5).

dataset	1.5-norm	2.5-norm	5-norm	10-norm	∞-norm	ELC	Banzhaf	Beta Shapley
GPSP	13.50	13.78	<u>14.11</u>	**14.17**	13.85	13.57	13.65	13.98
FOTP	35.76	37.44	<u>38.60</u>	**38.75**	36.87	35.81	36.82	37.76
wave_energy	17.11	<u>17.59</u>	**17.69**	17.29	16.35	16.69	17.32	17.29
jannis	8.13	15.78	<u>17.42</u>	**17.48**	15.45	15.56	13.75	14.09
spambase	28.29	<u>28.40</u>	28.39	28.14	27.40	27.53	**28.45**	28.34
superconduct	24.91	25.12	25.16	24.97	23.69	**26.34**	25.14	<u>25.43</u>
letter	0.56	1.01	1.14	<u>1.23</u>	1.09	**1.26**	0.86	0.43
pendigits	3.01	<u>3.15</u>	**3.16**	3.12	3.04	3.11	3.12	2.90
EES	**4.96**	<u>4.93</u>	4.87	4.68	4.49	4.12	4.92	4.89
WQW	0.34	0.47	<u>0.55</u>	**0.56**	0.46	0.42	0.46	0.23
elevators	8.43	8.54	8.63	8.65	8.59	**8.84**	8.47	<u>8.74</u>
credit	**3.40**	3.38	3.22	3.01	2.98	3.17	3.38	<u>3.39</u>

over 30 different instances of $\mathbf{x}$. For classification datasets, f outputs the logit of a randomly selected class.

Task. The task is to rank players, where performance is evaluated using the inclusion AUC metric defined in problem (3). A higher value indicates better performance. According to Lemma 3.3, the player ranking π induced by a specific contribution vector $\boldsymbol{\phi}$ satisfies that $\phi_{\pi(j)} \geq \phi_{\pi(j+1)}$ for every $1 \leq j < n$. Intuitively, $U_f^{\mathbf{x}}$ is supposed to increase sharply after including the players that contribute the most.

Datasets. We employ 12 datasets in total, which are summarized in Table 2. In particular, the datasets we selected are equally divided into two group. Each dataset in the first group contains fewer than 20 features, whereas those in the second group contain more than 30 features. For decision trees trained using the first group of datasets, we solve the exact minimization problem for each considered nonlinear attribution method. For the other group of datasets, the budget B in Algorithm 1 is set to $\#\text{players} \times 1,000$. For the two regression datasets, we scale the predicted values for better presentation.

Baselines. We compare the selected nonlinear attribution methods to Beta Shapley values [Kwon and Zou, 2022a]. The considered candidates include Beta(α, β) with

$$(\alpha, \beta) \in \{(16,1),(8,1),(4,1),(2,1),(1,1),\\ (1,2),(1,4),(1,8),(1,16),(1,32)\}, \tag{5}$$

which were used in [Kwon and Zou, 2022b]. Meanwhile, we also compare with another well-known semi-value, the Banzhaf value [Banzhaf III, 1965]. Therefore, a total of 11 linear attribution methods are used.

For reproducibility, we fix the random seed for training GBTs and selecting logits. When performing approximation, we average the results over five different random seeds and report the mean along with the corresponding standard deviation. More experimental details and statistical results can be found in Appendix B.

5.1 EMPIRICAL OBSERVATIONS

Results for utility functions containing fewer than 20 players are shown in Figure 2, while those for datasets with more than 30 players are presented in Figure 1. For clarity, we plot only the inclusion curves of the best linear and nonlinear candidates that maximize the inclusion AUC. In Appendix B, we also plot the inclusion curves for all employed nonlinear attribution methods. The average inclusion AUC scores are reported in Table 1. We make the following observations: (i) nonlinear attribution methods tend to outperform linear attribution methods under the inclusion AUC metric; and (ii) no single attribution method consistently outperforms others across all datasets.

6 CONCLUSION

While the Shapley value is widely adopted for its axiomatic properties, we demonstrate that its linearity in utility functions can lead to unreliable/undesirable player ranking under the inclusion AUC metric. Moreover, the efficiency axiom has also been empirically proved unfavorable. Accordingly, we introduce a class of attribution methods by relaxing the linearity and efficiency axioms while preserving other desirable axioms such as consistency, equal treatment, monotonicity, and translation invariance. In particular, these introduced nonlinear attribution methods are based on faithful additive approximation of utility functions. We view our work as the first step to establish nonlinear axiomatic attribution methods through the lens of additive approximation.

Acknowledgements

This research is supported by the National Research Foundation Singapore and the Singapore Ministry of Digital Development and Innovation, National AI Group under the AI Visiting Professorship Programme (award number AIVP-2024-001). YY gratefully acknowledges NSERC and CIFAR for funding support.

REFERENCES

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1): 42–60, 2018.

Javier Arin, Jeroen Kuipers, and Dries Vermeulen. An axiomatic approach to egalitarianism in TU-games. *International Journal of Game Theory*, 37(4):565–580, 2008.

John F Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19(2):317–343, 1965.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2): e2304406120, 2024.

E Blum and Werner Oettli. Direct proof of the existence theorem for quadratic programming. *Operations Research*, 20(1):165–167, 1972.

James P Bridge, Sean B Holden, and Lawrence C Paulson. Machine learning for first-order theorem proving: Learning to select a good heuristic. *Journal of automated reasoning*, 53:141–172, 2014.

Abraham Charnes, Boaz Golany, Michael S. Keane, and John J. Rousseau. Extremal principle solutions of games in characteristic function form: Core, Chebychev and Shapley value generalizations. In *Econometrics of Planning and Efficiency*, pages 123–133, 1988.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807, 2024.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

Ian Connick Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate Shapley values with vision transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Pradeep Dubey, Abraham Neyman, and Robert James Weber. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981.

Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings 17*, pages 535–546. Springer, 2015.

Peter W Frey and David J Slate. Letter recognition using holland-style adaptive classifiers. *Machine learning*, 6: 161–182, 1991.

Ian Gemp, Marc Lanctot, Luke Marris, Yiran Mao, Edgar Duéñez-Guzmán, Sarah Perrin, Andras Gyorgy, Romuald Elie, Georgios Piliouras, Michael Kaisers, et al. Approximating the core via iterative coalition sampling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 669–678, 2024.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*, 12(11):1610–1623, 2019a.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the Shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2019b.

Adam Karczmarz, Tomasz Michalak, Anish Mukherjee, Piotr Sankowski, and Piotr Wygocki. Improved feature importance computation for tree models based on the Banzhaf value. In *Uncertainty in Artificial Intelligence*, pages 969–979. PMLR, 2022.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pages 5491–5500. PMLR, 2020.

Yongchan Kwon and James Y Zou. Beta Shapley: A unified and noise-reduced data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8780–8802, 2022a.

Yongchan Kwon and James Y Zou. WeightedSHAP: Analyzing and improving Shapley based feature attributions. *Advances in Neural Information Processing Systems*, 35: 34363–34376, 2022b.

Weida Li and Yaoliang Yu. Robust data valuation with weighted Banzhaf values. *Advances in Neural Information Processing Systems*, 36, 2023.

Weida Li and Yaoliang Yu. One sample fits all: Approximating all probabilistic values simultaneously and efficiently. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Weida Li, Yaoliang Yu, and Bryan Kian Hsiang Low. TreeGrad-Ranker: Feature ranking via $O(L)$-time gradients for decision trees. In *The Fourteenth International Conference on Learning Representations*, 2026.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Renata CB Madeo, Clodoaldo AM Lima, and Sarajane M Peres. Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 46–52, 2013.

Jean-Luc Marichal and Pierre Mathonet. Weighted Banzhaf power and interaction indexes through weighted approximations of games. *European journal of operational research*, 211(2):352–358, 2011.

Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. Beyond treeSHAP: Efficient computation of any-order Shapley interactions for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14388–14396, 2024.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *The British Machine Vision Conference (BMVC)*, 2018.

Oliver Roesler. EEG Eye State. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C57G7J.

Lloyd S Shapley. A value for n-person games. *Annals of Mathematics Studies*, 28:307–317, 1953.

Lloyd S Shapley. Cores of convex games. *International journal of game theory*, 1:11–26, 1971.

Jiachen T Wang and Ruoxi Jia. Data Banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6388–6421. PMLR, 2023.

Jiachen T. Wang, Tianji Yang, James Zou, Yongchan Kwon, and Ruoxi Jia. Rethinking Data Shapley for data selection tasks: Misleads and merits. In *Forty-first International Conference on Machine Learning*, 2024.

Robert James Weber. *Probabilistic values for games*, page 101–120. Cambridge University Press, 1988.

Tom Yan and Ariel D Procaccia. If you like Shapley then you'll love the core. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5751–5759, 2021.

Koji Yokote, Yukihiko Funaki, and Yoshio Kamijo. A new basis and the Shapley value. *Mathematical social sciences*, 80:21–24, 2016.

Peng Yu, Chao Xu, Albert Bifet, and Jesse Read. Linear TreeShap. *Advances in Neural Information Processing Systems*, 35:25818–25828, 2022.

Nonlinear Axiomatic Attribution for Cooperative Games (Supplementary Material)

Weida Li[1] **Zhuanghua Liu**[1] **Yaoliang Yu**[2,3] **Bryan Kian Hsiang Low**[1]

[1] Department of Computer Science, National University of Singapore, Republic of Singapore
[2] School of Computer Science, University of Waterloo, Canada
[3] Vector Institute, Canada

A MISSING PROOFS

Theorem 3.1. *The egalitarian least core ϕ^{ELC} is nonlinear and satisfies the axioms of consistency, equal treatment, monotonicity, efficiency, and translation invariance.*

Proof. **Efficiency**. This is imposed by the constraint $\sum_{i\in[n]} x_i = U_n([n]) - U_n(\emptyset)$ that appears in the problem (1).

Translation Invariance. This is clear as it is $\{U_n(S_1) - U_n(S_2)\}$ that matters in the problem (1).

Equal Treatment. Suppose there exist $i, j \in [n]$ such that $U_n(S \cup \{i\}) = U_n(S \cup \{j\})$ for every $S \subseteq [n] \setminus \{i, j\}$. For the sake of contradiction, assume $\phi_i^{\mathrm{ELC}}(U_n) > \phi_j^{\mathrm{ELC}}(U_n)$. Define a contribution vector $\boldsymbol{\phi} \in \mathbb{R}^n$ by letting

$$\phi_k \coloneqq \begin{cases} \phi_k^{\mathrm{ELC}}(U_n), & k \in [n] \setminus \{i, j\} \\ \frac{\phi_i^{\mathrm{ELC}}(U_n) + \phi_j^{\mathrm{ELC}}(U_n)}{2}, & \text{otherwise.} \end{cases}$$

Let

$$F(S) \coloneqq U_n(S \cup \{i\}) - U_n(\emptyset) - \sum_{k\in S} \phi_k^{\mathrm{ELC}}(U_n).$$

Since

$$F(S) - \phi_i \le \max(F(S) - \phi_i^{\mathrm{ELC}}(U_n),\ F(S) - \phi_j^{\mathrm{ELC}}(U_n),\ 0)$$

for every $S \subseteq [n] \setminus \{i, j\}$, $\boldsymbol{\phi}$ with $e = \max_{S\subseteq[n]} U_n(S) - \sum_{i\in S} \phi_i^{\mathrm{ELC}}(U_n)$ is another optimal solution to the problem (1) . However,

$$2 \times \left(\frac{\phi_i^{\mathrm{ELC}}(U_n) + \phi_j^{\mathrm{ELC}}(U_n)}{2}\right)^2 < \phi_i^{\mathrm{ELC}}(U_n)^2 + \phi_j^{\mathrm{ELC}}(U_n)^2,$$

a contradiction to that $\phi^{\mathrm{ELC}}(U_n)$ is the least-norm optimal solution.

Consistency. Suppose there exists $C \in \mathbb{R}$ such that $U_{n+1}(S) = U_n(S \cap [n]) + C \cdot \mathbb{1}_{n+1\in S}$ for every $S \subseteq [n+1]$. First, we prove that $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) \ge C$. For convenience, we write

$$d(S) \coloneqq U_{n+1}(S) - U_{n+1}(\emptyset) - \sum_{i\in S} \phi_i^{\mathrm{ELC}}(U_{n+1}) \ \text{ for every } \ S \subseteq [n+1]$$
$$\text{and } \ e = \max_{S\subseteq[n]} d(S) \ge C - U_{n+1}^{\mathrm{ELC}}(U_{n+1}) > 0.$$

For the sake of contradiction, assume $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) < C$. Then, we have

$$d(S) < d(S) + (C - U_{n+1}^{\mathrm{ELC}}(U_{n+1})) = d(S \cup \{n+1\}) \ \text{ for every } \ S \subseteq [n].$$

It indicates that

$$d(S) = e \implies n+1 \in S.$$

Define a contribution $\boldsymbol{\phi} \in \mathbb{R}^{n+1}$ by letting

$$\phi_k \coloneqq \begin{cases} \phi_k^{\mathrm{ELC}}(U_{n+1}) + \frac{C - \phi_{n+1}^{\mathrm{ELC}}(U_{n+1})}{3}, & k = n+1, \\ \phi_k^{\mathrm{ELC}}(U_{n+1}) - \frac{C - \phi_{n+1}^{\mathrm{ELC}}(U_{n+1})}{3n}, & \text{otherwise.} \end{cases}$$

Then, $\boldsymbol{\phi}$ satisfies the efficiency constraint, and for every $S \subseteq [n]$

$$\begin{aligned} U_{n+1}(S) - U_{n+1}(\emptyset) - \sum_{i \in S} \phi_i &\le d_{n+1}(S) + \frac{C - \phi_{n+1}^{\mathrm{ELC}}(U_{n+1})}{3} \\ &< d_{n+1}(S \cup \{n+1\}) - \frac{C - \phi_{n+1}^{\mathrm{ELC}}(U_{n+1})}{3} \\ &\le U_{n+1}(S \cup \{n+1\}) - U_{n+1}(\emptyset) - \sum_{i \in S \cup \{n+1\}} \phi_i. \end{aligned}$$

In particular,

$$U_{n+1}(S \cup \{n+1\}) - U_{n+1}(\emptyset) - \sum_{i \in S \cup \{n+1\}} \phi_i < d(S \cup \{n+1\})$$

for every $S \subsetneq [n]$, which contradicts the optimality of $(\phi^{\mathrm{ELC}}(U_{n+1}),\, e)$. Therefore, we have $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) \ge C$.

To have $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) \le C$, the same argument applies by noticing that

$$e \ge d([n]) > d([n+1]) = 0$$

if we assume $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) > C$. As a result, we have $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) = C$. With a moment's thought, $\phi_{n+1}^{\mathrm{ELC}}(U_{n+1}) = C$ implies

$$\phi_k^{\mathrm{ELC}}(U_{n+1}) = \phi_k^{\mathrm{ELC}}(U_n) \ \text{ for every } \ k \in [n].$$

Monotonicity. Suppose there exists $j \in [n]$ such that $U_n(S \cup \{j\}) \ge U_n(S)$ for every $S \subseteq [n] \setminus \{j\}$. Assume, for the sake of contradiction, that $\phi_j^{\mathrm{ELC}} < 0$. Then,

$$r \coloneqq \max_{S \subseteq [n]} U_n(S) - U_n(\emptyset) - \sum_{i \in S} \phi_i^{\mathrm{ELC}}(U_n) \ge U_n(\{j\}) - U_n(\emptyset) - \phi_j^{\mathrm{ELC}}(U_n) > 0,$$

and, for every $S \subseteq [n] \setminus \{j\}$,

$$U_n(S) - U_n(\emptyset) - \sum_{i \in S} \phi_i^{\mathrm{ELC}}(U_n) < U_n(S \cup \{j\}) - U_n(\emptyset) - \sum_{i \in S \cup \{j\}} \phi_i^{\mathrm{ELC}}(U_n),$$

which indicates

$$U_n(S) - U_n(\emptyset) - \sum_{i \in S} \phi_i^{\mathrm{ELC}}(U_n) = r \implies j \in S.$$

Select $\delta > 0$ such that

$$\delta < \min_{S \subseteq [n] \setminus \{j\}} U_n(S \cup \{j\}) - U_n(S) - \phi_j^{\mathrm{ELC}}(U_n).$$

Define $\boldsymbol{\phi} \in \mathbb{R}^n$ by letting

$$\phi_k \coloneqq \begin{cases} \phi_k^{\mathrm{ELC}}(U_n) + \frac{\delta}{3}, & k = j \\ \phi_k^{\mathrm{ELC}}(U_n) - \frac{\delta}{3(n-1)}, & \text{otherwise.} \end{cases}$$

Observe that $\boldsymbol{\phi}$ satisfies the efficiency constraint, and for every $S \subseteq [n] \setminus \{j\}$

$$\begin{aligned} U_n(S) - U_n(\emptyset) - \sum_{i \in S} \phi_i &\le U_n(S) - U_n(\emptyset) - \sum_{i \in S} \phi_i^{\mathrm{ELC}}(U_n) + \frac{\delta}{3} \\ &< U_n(S \cup \{j\}) - U_n(\emptyset) - \sum_{i \in S \cup \{j\}} \phi_i^{\mathrm{ELC}}(U_n) - \frac{\delta}{3} \\ &\le U_n(S \cup \{j\}) - U_n(\emptyset) - \sum_{i \in S \cup \{j\}} \phi_i. \end{aligned}$$

Specifically,

$$U_n(S\cup\{j\})-U_n(\emptyset)-\sum_{S\cup\{j\}}\phi_i < U_n(S\cup\{j\})-U_n(\emptyset)-\sum_{S\cup\{j\}}\phi_i^{\mathrm{ELC}}(U_n)$$

for every $S\subsetneq[n]\setminus\{j\}$, contradicting the optimality of $(\phi^{\mathrm{ELC}}(U_n),\, r)$. Consequently, we have $\phi_j^{\mathrm{ELC}}(U_n)\geq 0$.

On the other hand, suppose there exists $j\in[n]$ such that $U_n(S\cup\{j\})\leq U_n(S)$ for every $S\subseteq[n]\setminus\{j\}$. Then, $\phi_j^{\mathrm{ELC}}(U_n)\leq 0$ can be proved similarly by noting that

$$r\geq U_n([n]\setminus\{j\})-U_n(\emptyset)-\sum_{i\in[n]\setminus\{j\}}\phi_i^{\mathrm{ELC}}(U_n) > U_n([n])-U_n(\emptyset)-\sum_{i\in[n]}\phi_i^{\mathrm{ELC}}(U_n)=0$$

if we assume $\phi_j^{\mathrm{ELC}}(U_n)>0$.

Non-Linearity. Define $U_2\in\mathcal{G}_2$ by letting

$$U_2(\emptyset)=U_2(\{1,2\})=0,\ U_2(\{1\})=-1,\ \text{and}\ U_2(\{2\})=-2.$$

Then, it is clear that $\phi^{\mathrm{ELC}}(U_2)=(0,0)^\top$. If ϕ^{ELC} is linear, it would reduce to the Shapley value [Shapley, 1953], as we have proved that ϕ^{ELC} satisfies the axioms of equal treatment, consistency, and efficiency. However, the Shapley value of U_2 is $(\frac{1}{2},-\frac{1}{2})^\top$, which indicates that ϕ^{ELC} must be non-linear. □

Theorem 4.1. *Let Π_n be the set of all permutations of $[n]$ with $n\geq 3$, and let ϕ^{Shap} denote the Shapley value. Then, Π_n can be partitioned into $\{\Pi_{n,j}\}_{j\in\mathcal{J}}$ with $|\mathcal{J}|\geq 2$ such that, for every $\Pi_{n,j}$ and every $U_n\in\mathcal{G}_n$, there exists a perturbation $U_n'\in\mathcal{G}_n$ satisfying*

$$\phi^{\mathrm{Shap}}(U_n)=\phi^{\mathrm{Shap}}(U_n+U_n'),$$
$$\textit{yet}\ \ \arg\max_{\pi\in\Pi_n}\mathrm{AUC}(\pi;U_n+U_n')\subseteq\Pi_{n,j}.$$

Proof. According to Yokote et al. [2016], a basis of the null space of the Shapley operator is $\{U_T\}_{T\in\mathcal{N}}$ where $\mathcal{N}:=\{S\subseteq[n]\mid |S|\neq 1\}$. Specifically,

$$U_\emptyset(S)=1\ \text{ for every }\ S\subseteq[n],$$
$$\text{and }\ U_T(S)=\begin{cases}1, & |T\cap S|=1,\\ 0, & \text{otherwise.}\end{cases}$$

Let $2^{[n]}$ and $\mathcal{N}$ be ordered so that every U_T can be written as a vector $\mathbf{a}_T\in\{0,1\}^{2^n}$, all of which constitute a matrix $\mathbf{A}\in\{0,1\}^{2^n\times|\mathcal{N}|}$. Then, every utility function U_n' in this null space can be expressed as $U_n'=\mathbf{A}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}\in\mathbb{R}^{|\mathcal{N}|}$.

Given a permutation $\pi\in\Pi_n$, we write

$$S_k(\pi)=\{\pi_1,\pi_2,\ldots,\pi_k\}\ \text{ for every } 0\leq k\leq n.$$

For each permutation $\pi\in\Pi_n$, it can be represented by a vector $\mathbf{p}_\pi\in\{0,1\}^{2^n}$ whose S-th entry is 1 if $S\in\{S_k(\pi)\colon 0\leq k\leq n\}$ and 0 otherwise. Then, we have

$$\mathrm{AUC}(\pi;U_n')=\mathbf{p}_\pi^\top\mathbf{A}\boldsymbol{\alpha}-U_n'(\emptyset).$$

For convenience, we write $\mathbf{d}_\pi^\top=\mathbf{p}_\pi^\top\mathbf{A}$ for every $\pi\in\Pi_n$. Next, we prove that every $\mathbf{d}_\pi^\top$ can be re-ordered to the same vector. Note that this is equivalent to ordering $\mathcal{N}$ according to the given π. Without loss of generality, we demonstrate this ordering process by using $n=4$ as an example.

For subsets that include π_1, their order is generated by

$$\{\pi_1\}$$
$$\{\pi_1\},\ \{\pi_1,\pi_4\},$$
$$\{\pi_1\},\ \{\pi_1,\pi_4\},\ \{\pi_1,\pi_3\},\ \{\pi_1,\pi_4,\pi_3\}$$
$$\{\pi_1\},\ \{\pi_1,\pi_4\},\ \{\pi_1,\pi_3\},\ \{\pi_1,\pi_4,\pi_3\},\ \{\pi_1,\pi_2\},\ \{\pi_1,\pi_4,\pi_2\},\ \{\pi_1,\pi_3,\pi_2\},\ \{\pi_1,\pi_4,\pi_3,\pi_2\}.$$

Following these ordered subsets, the subsets that include π_2 but not π_1 are ordered as

$$\{\pi_2\}$$
$$\{\pi_2\},\ \{\pi_2,\pi_4\}$$
$$\{\pi_2\},\ \{\pi_2,\pi_4\},\ \{\pi_2,\pi_3\},\ \{\pi_2,\pi_4,\pi_3\}.$$

Finally, it is

$$\{\pi_3\}$$
$$\{\pi_3\},\ \{\pi_3,\pi_4\}.$$

The empty set is appended to these ordered non-empty subsets. The order is finalized by removing all the singleton subsets. Precisely, the π-specific order for $\mathcal{N}$ is

$$\{\pi_1,\pi_4\},\ \{\pi_1,\pi_3\},\ \{\pi_1,\pi_4,\pi_3\},\ \{\pi_1,\pi_2\},\ \{\pi_1,\pi_4,\pi_2\},\ \{\pi_1,\pi_3,\pi_2\},\ \{\pi_1,\pi_4,\pi_3,\pi_2\},$$
$$\{\pi_2,\pi_4\},\ \{\pi_2,\pi_3\},\ \{\pi_2,\pi_4,\pi_3\},\ \{\pi_3,\pi_4\},\ \emptyset.$$

One can verify that, using this order for $\mathcal{N}$, $\mathbf{d}_\pi^\top = (3,2,2,1,1,1,1,2,1,1,1,5)$. In other words, for $1 \le i \le 2^n - 1$, the i-th entry of $\mathbf{d}_\pi$ is equal to the number of the occurrences of the corresponding subset in the ordering process, whereas the last entry is equal to $n+1$.

From now on, assume the order of $\mathcal{N}$ is fixed. Consequently, we can partition Π_n into $\{\Pi_{n,j}\}_j$ such that, for every $\pi,\pi' \in \Pi_{n,j}$, $\mathbf{d}_\pi = \mathbf{d}_{\pi'}$. From the ordering process, it is clear that $|\{\Pi_{n,j}\}_j| \ge 2$ if $n \ge 3$.

For each $\Pi_{n,j}$, let $U_n' = \mathbf{A}\mathbf{d}_\pi$ where $\pi \in \Pi_{n,j}$. It is clear that $\mathrm{AUC}(\pi; U_n') > \mathrm{AUC}(\pi'; U_n')$ for every $\pi' \notin \Pi_{n,j}$. By choosing an appropriate scalar $c > 0$, we have the result by setting $U_n' = c \cdot \mathbf{A}\mathbf{d}_\pi$.

□

Theorem 4.3. *For $p \in (1,\infty) \cup \{\infty\}$, the p-norm attribution method ϕ^p satisfies the axioms of consistency, equal treatment, monotonicity and translation invariance. When $n > 2$, ϕ^p is linear if and only if $p = 2$.*

Proof. We divide our proof into two parts. The first part is to demonstrate the properties of ϕ^∞, whereas the second part goes for $p \in (1,\infty)$. Obviously, ϕ^p with $p \in (0,\infty) \cup \{\infty\}$ satisfies the axiom of translation invariance due to the bias variable b.

For convenience, we write

$$d(S,\mathbf{x};U_n) \coloneqq U_n(S) - b_p(U_n) - \sum_{i\in S} x_i,$$
$$\text{and}\ \ e_p(U_n) \coloneqq \min_{\mathbf{x}\in\mathbb{R}^n, b\in\mathbb{R}} \|\mathbf{A}\mathbf{x} + b\cdot\mathbf{1}_{2^n} - \mathbf{V}(U_n)\|_p.$$

where $b_p(U_n)$ is the optimal value of b in then minimization problem with $\mathbf{x} = \phi^\infty(U_n)$.

Note that

$$e_\infty(U_n) = \frac{\max_{S\subseteq[n]} d(S,\phi^\infty(U_n);U_n) - \min_{S\subseteq[n]} d(S,\phi^\infty(U_n);U_n)}{2}. \tag{6}$$

Besides, for $p \in (0,\infty)$, define $\psi_p : \mathbb{R} \to \mathbb{R}$ by letting

$$\psi_p(t) = \mathrm{sign}(t)\cdot|t|^{p-1}.$$

ϕ^∞ satisfies the axiom of consistency.

This is straightforward by observing that the bias variable b ensures that there exist two subsets $S,T \subseteq [n]$ such that

$$d(S,\phi^\infty(U_n);U_n) = e_\infty(U_n)\ \text{ and }\ d(T,\phi^\infty(U_n);U_n) = -e_\infty(U_n).$$

ϕ^∞ satisfies the axiom of monotonicity. Suppose there exists $j \in [n]$ such that $U_n(S\cup\{j\}) \ge U_n(S)$ for every $S \subseteq [n]\setminus\{j\}$. Assume, for the sake of contradiction, that $\phi_j^\infty(U_n) < 0$. Then, for every subset $\mathbf{S} \subseteq [n]\setminus\{j\}$

$$d(S,\phi^\infty(U_n);U_n) < d(S\cup\{j\},\phi^\infty(U_n);U_n).$$

It indicates that

$$d(S,\phi^\infty(U_n);U_n)=e_\infty(U_n)\implies j\in S\ \text{ and }\ d(S,\phi^\infty(U_n);U_n)=-e_\infty(U_n)\implies j\notin S.$$

Select

$$0<\delta<\min_{S\subseteq[n]\setminus\{j\}}d(S\cup\{j\},\phi^\infty(U_n);U_n)-\min_{S\subseteq[n]\setminus\{j\}}d(S,\phi^\infty(U_n);U_n).$$

Define $\boldsymbol{\phi}\in\mathbb{R}^n$ by letting

$$\phi_k\coloneqq\begin{cases}\phi_k^\infty(U_n), & k\neq j,\\ \phi_k^\infty(U_n)+\delta, & \text{otherwise.}\end{cases}$$

Clearly,

$$\min_{S\subseteq[n]\setminus\{j\}}d(S\cup\{j\},\boldsymbol{\phi};U_n)<\min_{S\subseteq[n]\setminus\{j\}}d(S\cup\{j\},\phi^\infty(U_n);U_n),$$

which, by Eq. (6), contradicts the optimality of $(\phi^\infty(U_n),\ b_\infty(U_n))$. Therefore, we must have $\phi_j^\infty(U_n)\geq 0$. The other case can be proved similarly.

ϕ^∞ **satisfies the axiom of equal treatment**. Suppose there exist $i,j\in[n]$ such that $U_n(S\cup\{i\})=U_n(S\cup\{j\})$ for every $S\subseteq[n]\setminus\{i,j\}$. For the sake of contradiction, assume that $\phi_i^\infty(U_n)>\phi^\infty(U_n)$. Define $\boldsymbol{\phi}\in\mathbb{R}^n$ by letting

$$\phi_k\coloneqq\begin{cases}\frac{\phi_i^\infty(U_n)+\phi_j^\infty(U_n)}{2}, & k\in\{i,j\},\\ \phi_k^\infty(U_n), & \text{otherwise.}\end{cases}$$

Observe that for every $S\subseteq[n]\setminus\{i,j\}$

$$|d(S\cup\{i\},\boldsymbol{\phi};U_n)|<\max\left(d(S\cup\{i\},\phi^\infty(U_n);U_n),\ d(S\cup\{j\},\phi^\infty(U_n);U_n)\right).$$

It means that the solution of $(\boldsymbol{\phi},b^*)$ is also

$$2\times\left(\frac{\phi_i^\infty(U_n)+\phi_j^\infty(U_n)}{2}\right)^2<\phi_i^\infty(U_n)^2+\phi_j^\infty(U_n)^2,$$

a contradiction to that $\phi^\infty(U_n)$ is the least-norm optimal solution.

ϕ^∞ **is not linear**. Assume, for the sake of contradiction, that ϕ^∞ is linear. Since ϕ^∞ satisfies the axioms of linearity, consistency, equal treatment, and monotonicity, according to Weber [1988], $\phi^\infty(U_n)$ can be expressed as

$$\phi_i^\infty(U_n)=\sum_{S\subseteq[n]\setminus\{i\}}q_{|S|+1}[U_n(S\cup\{i\})-U_n(S)]\tag{7}$$

where $\mathbf{q}\in\mathbb{R}^n$ is non-negative and satisfies that $\sum_{k=1}^n\binom{n-1}{k-1}q_k=1$. For $t\in[3,5]$, define $U_3^t\in\mathcal{G}_3$ by letting

$$U_3^t(\emptyset)=0,\ U_3^t(\{1\})=U_3^t(\{2,3\})=U_3^t(\{1,2,3\})=4,\ U_3^t(\{2\})=U_3^t(\{3\})=3,$$
$$\text{and }\ U_3^t(\{1,2\})=U_3^t(\{1,3\})=t.$$

Using the constraints for $\emptyset$ and $\{1\}$, we have

$$b_\infty(U_3^t)\in[-e_\infty(U_3^t),\ e_\infty(U_3^t)]\ \text{ and }\ b+\phi_1^\infty(U_3^t)\in[4-e_\infty(U_3^t),\ 4+e_\infty(U_3^t)],$$

which leads to

$$\phi_1^\infty(U_3^t)\in[4-2e_\infty(U_3^t),\ 4+2e_\infty(U_3^t)].\tag{8}$$

Using the constraint for $\{2,3\}$, there is

$$b^*+\phi_2^\infty(U_3^t)+\phi_3^\infty(U_3^t)\in[4-e_\infty(U_3^t),\ 4+e_\infty(U_3^t)]\tag{9}$$

Combining Eqs. (8) and (9) yields

$$b^*+\phi_1^\infty(U_3^t)+\phi_2^\infty(U_3^t)+\phi_3^\infty(U_3^t)\in[8-3e_\infty(U_3^t),\ 8+3e_\infty(U_3^t)].$$

On the other hand, from the constraint for $\{1,2,3\}$, there is

$$b^* + \phi_1^\infty(U_3^t) + \phi_2^\infty(U_3^t) + \phi_3^\infty(U_3^t) \in [4 - e_\infty(U_3^t),\ 4 + e_\infty(U_3^t)].$$

Solving $4 + e_\infty(U_3^t) = 8 - 3e_\infty(U_3^t)$ yields an lower bound for $e_\infty(U_3^t)$, i.e., $e_\infty(U_3^t) \geq 1$. If $e^* = 1$, following the derivation of Eq. (8), the lower bounds for $\phi_1^\infty(U_3^t)$, $\phi_2^\infty(U_3^t)$, and $\phi_3^\infty(U_3^t)$ are 2, 1, and 1, respectively. Then, one can verify that the solution of $b = 1$ and $\mathbf{x} = (2,1,1)$ is indeed optimal, leading to $e_\infty(U_3^t) = 1$. In other words, $\phi^\infty(U_3^t) = (2,1,1)$ for every $t \in [3,5]$.

Then, it indicates that

$$q_1 = 1,\ q_2 = q_3 = 0.$$

However, Eq. (7) produces $\phi^\infty(U_3^t) = (4,3,3)$, contradicting to the uniqueness of $\phi^\infty(U_3^t)$. Therefore, ϕ^∞ must be nonlinear.

ϕ^p **with** $p \in (1,\infty)$ **satisfies the axiom of consistency**. Suppose there exists $C \in \mathbb{R}$ such that $U_{n+1}(S) = U_n(S \cap [n]) + C \cdot \mathbb{1}_{n+1 \in S}$ for every $S \subseteq [n+1]$. Since the corresponding minimization problem is strictly convex, using the first order condition for optimality, we have

$$\begin{aligned} &\sum_{S \subseteq [n]} \psi_p(d(S, \phi^p(U_n); U_n)) = 0 \\ \text{and} \quad &\sum_{S \subseteq [n] \setminus \{i\}} \psi_p(d(S \cup \{i\}, \phi^p(U_n); U_n)) = 0 \ \text{ for every } \ i \in [n]. \end{aligned} \tag{10}$$

Define $\boldsymbol{\phi} \in \mathbb{R}^{n+1}$ by letting $\phi_k = \phi_k^p(U_n)$ if $k \in [n]$ and C otherwise. Define

$$d'(S, \mathbf{x}) \coloneqq U_{n+1}(S) - b_p(U_n) - \sum_{i \in S} x_i.$$

Verifying the first order condition for U_{n+1},

$$\begin{aligned} \sum_{S \subseteq [n+1]} \psi_p(d'(S, \boldsymbol{\phi})) = 2 \sum_{S \subseteq [n]} \psi_p(d(S, \phi^p(U_n); U_n)) = 0, \\ \sum_{S \subseteq [n+1] \setminus \{i\}} \psi_d(d'(S \cup \{i\}, \boldsymbol{\phi})) = 2 \sum_{S \subseteq [n] \setminus \{i\}} \psi_p(d(S \cup \{i\}, \phi^p(U_n); U_n)) = 0 \ \text{ for every } i \in [n], \\ \sum_{S \subseteq [n]} \psi_d(d'(S \cup \{n+1\}, \boldsymbol{\phi})) = \sum_{S \subseteq [n]} \psi_d(d(S, \phi^p(U_n); U_n)) = 0. \end{aligned}$$

ϕ^p **with** $p \in (1,\infty)$ **satisfies the axiom of monotonicity**. Suppose there exists $j \in [n]$ such that $U_n(S \cup \{j\}) \geq U_n(S)$ for every $S \subseteq [n] \setminus \{j\}$. Using Eq. (10), we have

$$\sum_{S \subseteq [n] \setminus \{j\}} \psi_d(d(S, \phi^p(U_n); U_n)) = \sum_{S \subseteq [n] \setminus \{j\}} \psi_d(d(S \cup \{j\}, \phi^p(U_n); U_n)) = 0 \tag{11}$$

Notice that the function ψ_d is strictly increasing. If $\phi_j^p(U_n) < 0$, for every $S \subseteq [n] \setminus \{j\}$, we have

$$d(S, \phi^p(U_n); U_n) < d(S \cup \{j\}, \phi^p(U_n); U_n).$$

Then,

$$\sum_{S \subseteq [n] \setminus \{j\}} \psi_d(d(S, \phi^p(U_n); U_n)) < \sum_{S \subseteq [n] \setminus \{j\}} \psi_d(d(S \cup \{j\}, \phi^p(U_n); U_n)),$$

which makes Eq. (11) impossible. Therefore, $\phi_j^p(U_n) \geq 0$. The other case can be proved similarly.

ϕ^p **with** $p \in (1,\infty)$ **satisfies the axiom of equal treatment**. Suppose there exist $i, j \in [n]$ such that $U_n(S \cup \{i\}) = U_n(S \cup \{j\})$ for every $S \subseteq [n] \setminus \{i,j\}$. Using Eq. (10), we have

$$\sum_{S \subseteq [n] \setminus \{i,j\}} \phi_p(d(S \cup \{i\}, \phi^p(U_n); U_n)) = \sum_{S \subseteq [n] \setminus \{i,j\}} \phi_p(d(S \cup \{j\}, \phi^p(U_n); U_n)).$$

Since the function ϕ_p is strictly increasing, this equality implies that $\phi_i^p(U_n) = \phi_j^p(U_n)$.

ϕ^p with $p \in (1, \infty)$ is linear if and only if $p = 2$. It is obvious that ϕ^2 is linear; therefore, we focus on proving the converse. Suppose ϕ^p with $p \in (1, \infty)$ is linear. Since ϕ^p satisfies the axioms of linearity, consistency, equal treatment, and monotonicity, according to Weber [1988], we have

$$\phi_i^p(U_n) = \sum_{S \subseteq [n] \subseteq \{i\}} \gamma_{|S|+1}^{[n]} [U_n(S \cup \{i\}) - U_n(S)] \text{ for every } i \in [n] \tag{12}$$

where $\boldsymbol{\gamma}^{[n]} \in \mathbb{R}^n$ is non-negative and satisfies that $\sum_{k=1}^n \binom{n-1}{k-1} \gamma_k^{[n]} = 1$. Define $U_2 \in \mathcal{G}_2$ by letting

$$U_2(\emptyset) = 0 \text{ and } U_2(\{1\}) = U_2(\{2\}) = U_2(\{1,2\}) = 1.$$

Then,

$$\phi_1^p(U_2) = \phi_2^p(U_2) = \gamma_1^{[2]}.$$

Using the first order condition, as described in Eq. (10), there is

$$\psi_p(-b_p(U_2)) = \psi_p(\gamma_1^{[2]} + b_p(U_2) - 1) \text{ and } \psi_p(1 - \gamma_1^{[2]} - b_p(U_2)) = \psi_p(2\gamma_1^{[2]} + b_p(U_2) - 1).$$

Since ψ_p is strictly increasing, it leads to

$$-b_p(U_2) = \gamma_1^{[2]} + b_p(U_2) - 1 \text{ and } 1 - \gamma_1^{[2]} - b_p(U_2) = 2\gamma_1^{[2]} + b_p(U_2) - 1,$$

which, combined with $\gamma_1^{[2]} + \gamma_2^{[2]} = 1$, yields

$$\gamma_1^{[2]} = \gamma_2^{[2]} = 0.5.$$

Then, employing the axiom of consistency, we obtain

$$\gamma_1^{[3]} + \gamma_2^{[3]} = \gamma_2^{[3]} + \gamma_3^{[3]} = 0.5. \tag{13}$$

It is worth pointing out that ϕ^p is linear for every $p \in (1, \infty)$ if $n = 2$, and thus we require $n \geq 3$. Define $U_3 \in \mathcal{G}_3$ by letting

$$U_3(\emptyset) = U_3(\{2\}) = U_2(\{1,2\}) = U_2(\{1,3\}) = 0,\ U_3(\{1\}) = U_3(\{1,2,3\}) = 1,$$
$$U_2(\{3\}) = -2\gamma_1^{[3]}, \text{ and } U_2(\{2,3\}) = 1 - 2\gamma_1^{[3]}.$$

Using the first order condition, as described in Eq. (10), there is

$$\begin{aligned}\psi_p(d(\{3\}, \phi^p(U_3); U_3)) + \psi_p(d(\{1,3\}, \phi^p(U_3); U_3)) + \psi_p(d(\{2,3\}, \phi^p(U_3); U_3)) \\ + \psi_p(d(\{1,2,3\}, \phi^p(U_3); U_3)) = 0.\end{aligned} \tag{14}$$

Using Eqs. (12) and (13), one can verify that

$$d(\{3\}, \phi^p(U_3); U_3) = d(\{1,3\}, \phi^p(U_3); U_3) = -b_p(U_3) - \gamma_1^{[3]} - \gamma_3^{[3]},$$
$$\text{and } d(\{2,3\}, \phi^p(U_3); U_3) = d(\{1,2,3\}, \phi^p(U_3); U_3) = 1 - 3\gamma_3^{[3]} - b_p(U_3).$$

Thus, Eq. (14) can be simplified as

$$\psi_p(1 - 3\gamma_3^{[3]} - b_p(U_3)) = \psi_p(b_p(U_3) + \gamma_1^{[3]} + \gamma_3^{[3]}).$$

Since ψ_p is strictly increasing, it leads to

$$1 - 3\gamma_3^{[3]} - b_p(U_3) = b_p(U_3) + \gamma_1^{[3]} + \gamma_3^{[3]},$$

which yields

$$b_p(U_3) = 0.5 - 2\gamma_3^{[3]} - 0.5\gamma_1^{[3]}. \tag{15}$$

Using the first order condition again,

$$\begin{aligned}\psi_p(d(\emptyset, \phi^p(U_3); U_3)) + \psi_p(d(\{2\}, \phi^p(U_3); U_3)) + \psi_p(d(\{3\}, \phi^p(U_3); U_3)) \\ + \psi_p(d(\{2,3\}, \phi^p(U_3); U_3)) = 0.\end{aligned}$$

Using Eqs (12), (13) and (15),

$$d(\emptyset, \phi^p(U_3); U_3) = 0.75 - 2.5\gamma_2^{[3]},$$
$$d(\{2\}, \phi^p(U_3); U_3) = 0.25 - 1.5\gamma_2^{[3]},$$
$$d(\{3\}, \phi^p(U_3); U_3) = -0.5\gamma_2^{[3]} - 0.25,$$
$$\text{and } d(\{2,3\}, \phi^p(U_3); U_3) = 0.5\gamma_2^{[3]} + 0.25.$$

Since the function ψ_p is odd and strictly increasing, we have

$$0.75 - 2.5\gamma_2^{[3]} = -0.25 + 1.5\gamma_2^{[3]},$$

which, combined with Eq. (13), yields

$$\gamma_1^{[3]} = \gamma_2^{[3]} = \gamma_3^{[3]} = 0.25. \tag{16}$$

Define $U_3' \in \mathcal{G}_3$ by letting

$$U_3'(\{1\}) = U_3'(\{1,3\}) = U_3'(\{2,3\}) = U_3'(\{1,2,3\}) = U_3'(\{2\}) = 0,$$
$$U_3'(\{3\}) = U_3'(\emptyset) = -4, \text{ and } U_3'(\{1,2\}) = 8.$$

Using the first order condition,

$$\psi_p(d(\emptyset, \phi^p(U_3'); U_3)) + \psi_p(d(\{2\}, \phi^p(U_3'); U_3)) + \psi_p(d(\{3\}, \phi^p(U_3'); U_3))$$
$$+\psi_p(d(\{2,3\}, \phi^p(U_3'); U_3)) = 0,$$

which, combined with Eqs. (12) and (16), is equal to

$$\psi_p(-b_p(U_3') - 4) = \psi_p(b_p(U_3') + 2).$$

Since ψ_p is strictly increasing,

$$b_p(U_3') = -3. \tag{17}$$

Using the first order condition again,

$$\psi_p(d(\{3\}, \phi^p(U_3'); U_3)) + \psi_p(d(\{1,3\}, \phi^p(U_3'); U_3)) + \psi_p(d(\{2,3\}, \phi^p(U_3'); U_3))$$
$$+\psi_p(d(\{1,2,3\}, \phi^p(U_3'); U_3)) = 0,$$

which, combined with Eqs. (12), (16) and (17), is equal to

$$3 \cdot 1^{p-1} = 3\psi_p(1) = \psi_p(3) = 3^{p-1}.$$

Solving $3^{p-2} = 1$ yields $p = 2$. □

Theorem 4.5. *For every $\eta > 0$, let $\phi^\infty(U_n; \eta)$ denote an optimal solution of $\mathbf{x}$ to*

$$\underset{\mathbf{x}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \ \|\mathbf{A}\mathbf{x} + b \cdot \mathbf{1}_{2^n} - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}\|_2^2.$$

Then, we have $\phi^\infty(U_n; \eta) \to \phi^\infty(U_n)$ as $\eta \to 0$.

Proof. The minimization problem can be equivalently rewritten as

$$\underset{\mathbf{x}\in\mathbb{R}^n, b, e\in\mathbb{R}}{\text{minimize}} \ e + \eta \cdot \|\mathbf{x}\|_2^2$$
$$\text{subject to } \ |U_n(S) - b - \sum_{i\in S} x_i| \le e \ \text{ for every } S \subseteq [n].$$

Then, according to Blum and Oettli [1972], $\phi^\infty(U_n; \eta)$ exists. To prove $\phi^\infty(U_n; \eta) \to \phi^\infty(U_n)$ as $\eta \to 0$, it is sufficient to prove that for every sequence $(\phi^\infty(U_n; \eta_k))_{k=1}^\infty$ such that $\eta_k \to 0$ as $k \to \infty$, there is $\phi^\infty(U_n; \eta_k) \to \phi^\infty(U_n)$ as $k \to \infty$.

For convenience, we write

$$\mathbf{x}_\eta \coloneqq \phi^\infty(U_n; \eta),$$

and let b_η be the optimal solution of b associated with $\mathbf{x}_\eta$. Besides, write $\mathbf{x}_0 \coloneqq \phi^\infty(U_n)$, and define b_0 similarly. Additionally, we append $\mathbf{1}_{2^n}$ as the last column to $\mathbf{A}$, the result of which is denoted by $\mathbf{B}$, and write $\mathbf{y}_\eta \coloneqq (\mathbf{x}_\eta, b_\eta) \in \mathbb{R}^{n+1}$.

Observe that

$$\begin{aligned}\|\mathbf{A}\mathbf{x}_\eta + b_\eta \cdot \mathbf{1}_{2^n} - \mathbf{V}(U_n)\|_\infty &= \|\mathbf{B}\mathbf{y}_\eta - \mathbf{V}(U_n)\|_\infty \\ &\le \|\mathbf{B}\mathbf{y}_\eta - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}_\eta\|_2^2 \\ &\le \|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}_0\|_2^2.\end{aligned} \tag{18}$$

It indicates that $(\mathbf{y}_{\eta_k})_{k=1}^\infty$ is eventually bounded. Precisely,

$$\begin{aligned}\|\mathbf{y}_\eta\|_2 &\le \frac{1}{\sigma_{\min}(\mathbf{B})}\|\mathbf{B}\mathbf{y}_\eta\|_2 \\ &\le \frac{C_\infty}{\sigma_{\min}(\mathbf{B})}\|\mathbf{B}\mathbf{y}_\eta\|_\infty \\ &\le \frac{C_\infty}{\sigma_{\min}(\mathbf{B})}\left(\|\mathbf{B}\mathbf{y}_\eta - \mathbf{V}(U_n)\|_\infty + \|\mathbf{V}(U_n)\|_\infty\right) \\ &\le \frac{C_\infty}{\sigma_{\min}(\mathbf{B})}\left(\|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}_0\|_2^2 + \|\mathbf{V}(U_n)\|_\infty\right).\end{aligned}$$

For the first inequality, $\sigma_{\min}(\mathbf{B})$ denotes the smallest singular value of $\mathbf{B}$; $\sigma_{\min}(\mathbf{B}) > 0$ as $\mathbf{B}$ has full column rank.[1] The second inequality follows from the fact that all norms on $\mathbb{R}^n$ are equivalent.

Since the sequence $(\mathbf{y}_{\eta_k})_{k=1}^\infty$ is eventually bounded, it is valid to select some convergent subsequence. Therefore, suffice it to prove that every convergent subsequence of $(\mathbf{x}_{\eta_k})_{k=1}^\infty$ converges to the same point $\mathbf{x}_0$.[2]

For simplicity, we abuse the notation $(\mathbf{y}_{\eta_k})_{k=1}^\infty$ to refer to a convergent subsequence of $(\mathbf{y}_{\eta_k})_{k=1}^\infty$.[3] Thus, we have $\mathbf{y}_{\eta_k} \to \mathbf{y}'$ as $k \to \infty$ for some point $\mathbf{y}' \in \mathbb{R}^{n+1}$, the first n entries of which is denoted by $\mathbf{x}'$. Since

$$\|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty \le \|\mathbf{B}\mathbf{y}_\eta - \mathbf{V}(U_n)\|_\infty,$$

combining this with Eq. (18), we have

$$0 \le \|\mathbf{B}\mathbf{y}_\eta - \mathbf{V}(U_n)\|_\infty - \|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty \le \eta \cdot \|\mathbf{x}_0\|_2^2.$$

This suggests that $\|\mathbf{B}\mathbf{y}_{\eta_k} - \mathbf{V}(U_n)\|_\infty \to \|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty$ as $k \to \infty$. Since the function $\mathbf{y} \mapsto \|\mathbf{B}\mathbf{y} - \mathbf{V}(U_n)\|_\infty$ is continuous, we have

$$\|\mathbf{B}\mathbf{y}' - \mathbf{V}(U_n)\|_\infty = \|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty. \tag{19}$$

On the other hand, since $\mathbf{x} \mapsto \|\mathbf{x}\|_2^2$ is continuous, there is

$$\lim_{k\to\infty} \|\mathbf{x}_{\eta_k}\|_2^2 \to \|\mathbf{x}'\|_2^2.$$

Besides, note that

$$\begin{aligned}\|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}_\eta\|_2^2 &\le \|\mathbf{B}\mathbf{y}_\eta - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}_\eta\|_2^2 \\ &\le \|\mathbf{B}\mathbf{y}_0 - \mathbf{V}(U_n)\|_\infty + \eta \cdot \|\mathbf{x}_0\|_2^2,\end{aligned}$$

which implies

$$\|\mathbf{x}_\eta\|_2^2 \le \|\mathbf{x}_0\|_2^2.$$

Then,

$$\|\mathbf{x}'\|_2^2 \le \|\mathbf{x}_0\|_2^2. \tag{20}$$

Using Eqs. (19) and (20), and the uniqueness of $\mathbf{x}_0$, we obtain

$$\mathbf{x}' = \mathbf{x}_0.$$

□

[1] We have proved that $\mathbf{B}^\top\mathbf{B}$ is invertible in [Li and Yu, 2024, Theorem 2]

[2] This is equivalent to proving that $\liminf_{k\to\infty} a_k = \limsup_{k\to\infty} a_k$ for a sequence $(a_k)_{k=1}^\infty$.

[3] Given a subsequence $(\mathbf{x}_{\eta_{k_j}})_{j=1}^\infty$ that converges to $\mathbf{x}'$, it may happen that $b_{\eta_{k_j}}$ does not converge. Nevertheless, we can further select a subsequence of it such that the corresponding bias term converges, and $\mathbf{x}'$ is still the limit.

Theorem 4.6. *For every $\eta > 0$, let $\phi^{\mathrm{ELC}}(U_n;\eta)$ denote the unique optimal solution of $\mathbf{x}$ to*

$$\begin{aligned} &\underset{\mathbf{x}\in\mathbb{R}^n, e\in\mathbb{R}}{\text{minimize}} \ e + \eta\cdot\|\mathbf{x}\|_2^2 \\ &\textit{subject to } \ U_n(S) - A_n(S;\mathbf{x},U_n(\emptyset)) \le e \ \textit{ for every } S\subsetneq[n] \\ &\textit{and } \ A_n([n];\mathbf{x},U_n(\emptyset)) = U_n([n]). \end{aligned}$$

Then, we have $\phi^{\mathrm{ELC}}(U_n;\eta)\to\phi^{\mathrm{ELC}}(U_n)$ as $\eta\to 0$.

Proof. According to Blum and Oettli [1972], $\phi^{\mathrm{ELC}}(U_n;\eta)$ exists. For every $\eta>0$, let the produced optimal solution be denoted by $(\mathbf{x}_\eta, e_\eta)$, and similarly we define $(\mathbf{x}_0, e_0)$ as the optimal solution corresponding to the egalitarian least core. In other words, $\mathbf{x}_\eta = \phi^{\mathrm{ELC}}(U_n;\eta)$ and $\mathbf{x}_0 = \phi^{\mathrm{ELC}}(U_n)$. Then, suffice it to prove that for every sequence $(\mathbf{x}_{\eta_k})_{k=1}^n$ such that $\eta_k\to 0$ as $k\to\infty$, there is $\mathbf{x}_{\eta_k}\to\mathbf{x}_0$ as $k\to\infty$.

Notice that

$$e_0 \le e_\eta \le e_\eta + \eta\cdot\|\mathbf{x}_\eta\|_2^2 \le e_0 + \eta\cdot\|\mathbf{x}_0\|_2^2,$$

which yields

$$0 \le e_\eta - e_0 \le \eta\cdot\|\mathbf{x}_0\|_2^2. \tag{21}$$

Therefore, $e_{\eta_k}\to e_0$ as $k\to\infty$. On the other hand,

$$e_0 + \eta\cdot\|\mathbf{x}_\eta\|_2^2 \le e_\eta + \eta\cdot\|\mathbf{x}_\eta\|_2^2 \le e_0 + \eta\cdot\|\mathbf{x}_0\|_2^2,$$

which leads to

$$\|\mathbf{x}_\eta\|_2^2 \le \|\mathbf{x}_0\|_2^2.$$

Therefore, the sequence $(\mathbf{x}_{\eta_k})_{k=1}^\infty$ is eventually bounded, and thus we can select a convergent subsequence. In particular, there must be

$$\liminf_{k\to\infty}\|\mathbf{x}_{\eta_k}\|_2^2 = \limsup_{k\to\infty}\|\mathbf{x}_{\eta_k}\|_2^2 = \|\mathbf{x}_0\|_2^2.$$

Otherwise, there would exist a subsequence of $(\mathbf{x}_{\eta_k})_{k=1}^\infty$ converging to some point $\mathbf{x}'\in\mathbb{R}^n$ such that

$$\|\mathbf{x}'\|_2^2 = \liminf_{k\to\infty}\|\mathbf{x}_{\eta_k}\|_2^2 < \|\mathbf{x}_2\|_2^2,$$

which, together with Eq. (21), contradicts the uniqueness of $\mathbf{x}_0$. Note that $(\mathbf{x}', e_0)$ is a feasible solution, since the domain is closed. Consequently,

$$\lim_{k\to\infty}\|\mathbf{x}_{\eta_k}\|_2^2 = \|\mathbf{x}_0\|_2^2.$$

Then, every convergent subsequence of $(\mathbf{x}_{\eta_k})_{k=1}^\infty$ must converge to the same point $\mathbf{x}_0$, since $\mathbf{x}_0$ is the unique vector such that (i) $(e_0,\mathbf{x}_0)$ is feasible to the problem of least core, and (ii) for any feasible $(e_0,\mathbf{x})$, it holds that $\|\mathbf{x}_0\|_2^2 \le \|\mathbf{x}\|_2^2$. As a result, we have $\mathbf{x}_{\eta_k}\to\mathbf{x}_0$ as $k\to\infty$. □

B EXPERIMENTS

The datasets used are summarized in Table 2. More experimental results on the inclusion curves are presented in Figures 3 and 4.

Statistical comparison. As shown in Figures 5 and 6, we also compare these methods statistically. For the value associated with methods A (row) and B (column), it is computed as

$$\frac{\sum_{i=1}^m[\mathbb{1}\{\mathrm{AUC}_i^A > \mathrm{AUC}_i^B\} - \mathbb{1}\{\mathrm{AUC}_i^A < \mathrm{AUC}_i^B\}]}{m}$$

where m is the total number of instances used and AUC_i^A is the AUC achieved by method A on the utility function created using the i-th instance. Therefore, the larger it is, the better method A is.

Table 2: Summary of the used datasets.

Dataset	#Instances	#Features	Source	Task	#Classes
GPSP [Madeo et al., 2013]	9,873	32	https://openml.org/d/4538	classification	5
FOTP [Bridge et al., 2014]	6,118	51	https://openml.org/d/1475	classification	6
wave_energy	72,000	48	https://openml.org/d/44975	regression	-
jannis	83,733	54	https://openml.org/d/41168	classification	4
spambase	4,601	57	https://openml.org/d/44	classification	2
superconduct	21,263	81	https://openml.org/d/43174	regression	-
letter [Frey and Slate, 1991]	20,000	16	https://openml.org/d/6	classification	26
pendigits	10,992	16	https://openml.org/d/32	classification	10
EES [Roesler, 2013]	14,980	14	https://openml.org/d/1471	classification	2
WQW [Cortez et al., 2009]	4,898	11	https://openml.org/d/40498	classification	7
elevators	16,599	18	https://openml.org/d/846	classification	2
credit	16,714	10	https://openml.org/d/44089	classification	2

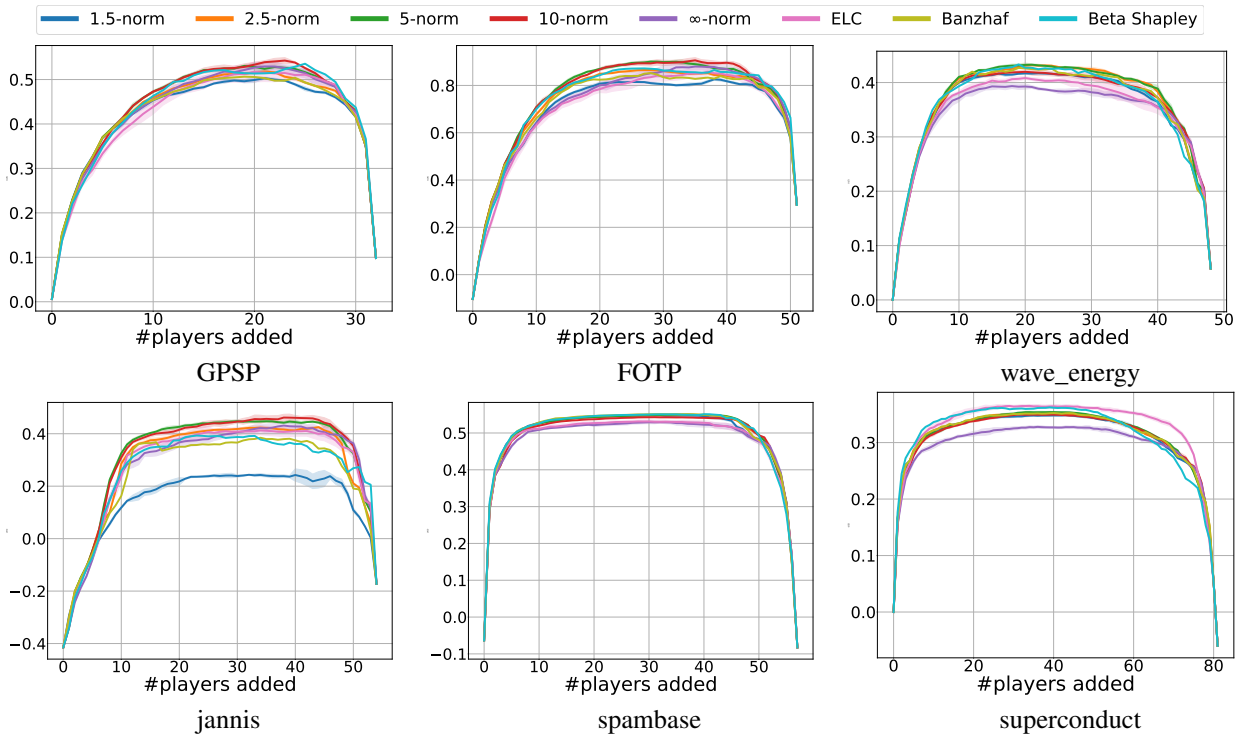

Figure 3: Comparison of attribution methods on six datasets for player ranking. A larger area under the curve indicates better performance. Since all utility functions contain more than 30 players, nonlinear methods are approximated by sampling $\#\text{players} \times 1,000$ subsets, with mean and standard deviation reported over five random seeds. Linear methods are computed exactly. Beta Shapley achieves the best inclusion AUC among the 10 candidates.

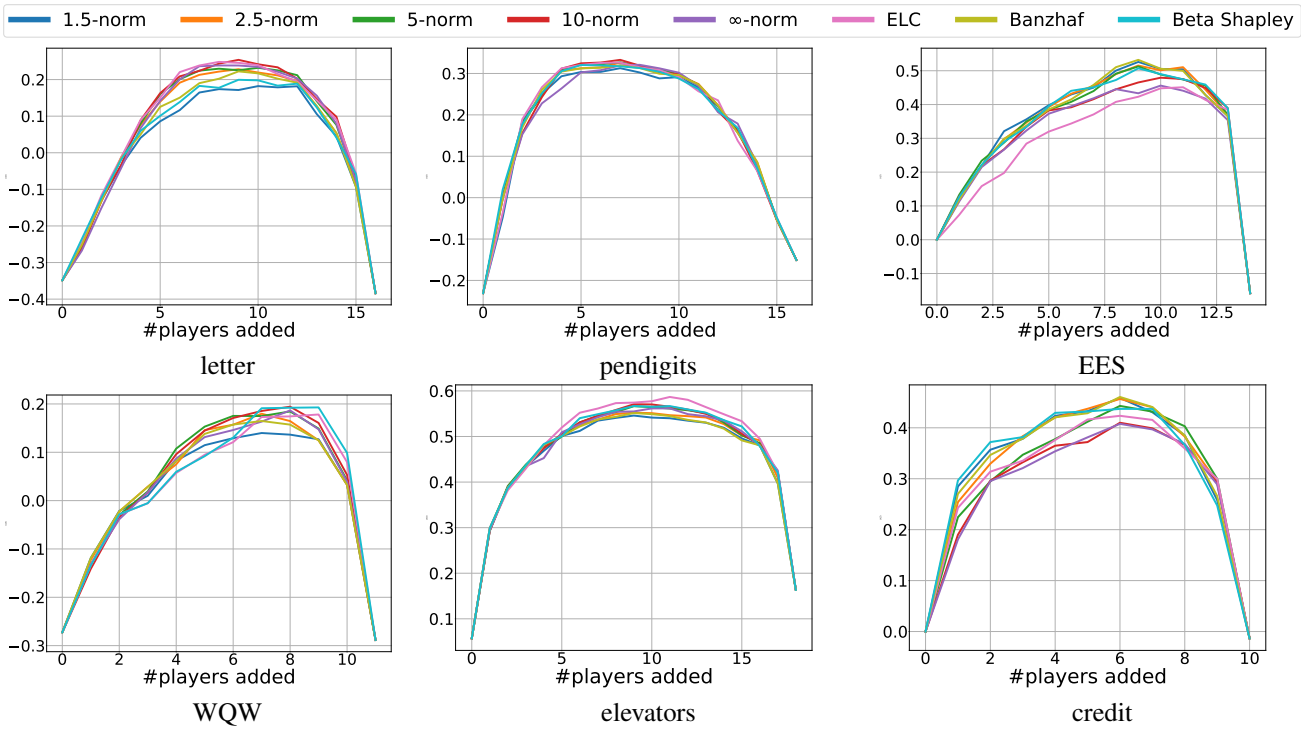

Figure 4: Comparison of attribution methods on six datasets for player ranking. A larger area under the curve indicates better performance. Since all utility functions have fewer than 20 players, all subsets are enumerated exactly for all methods. Beta Shapley achieves the best inclusion AUC among the 10 candidates.

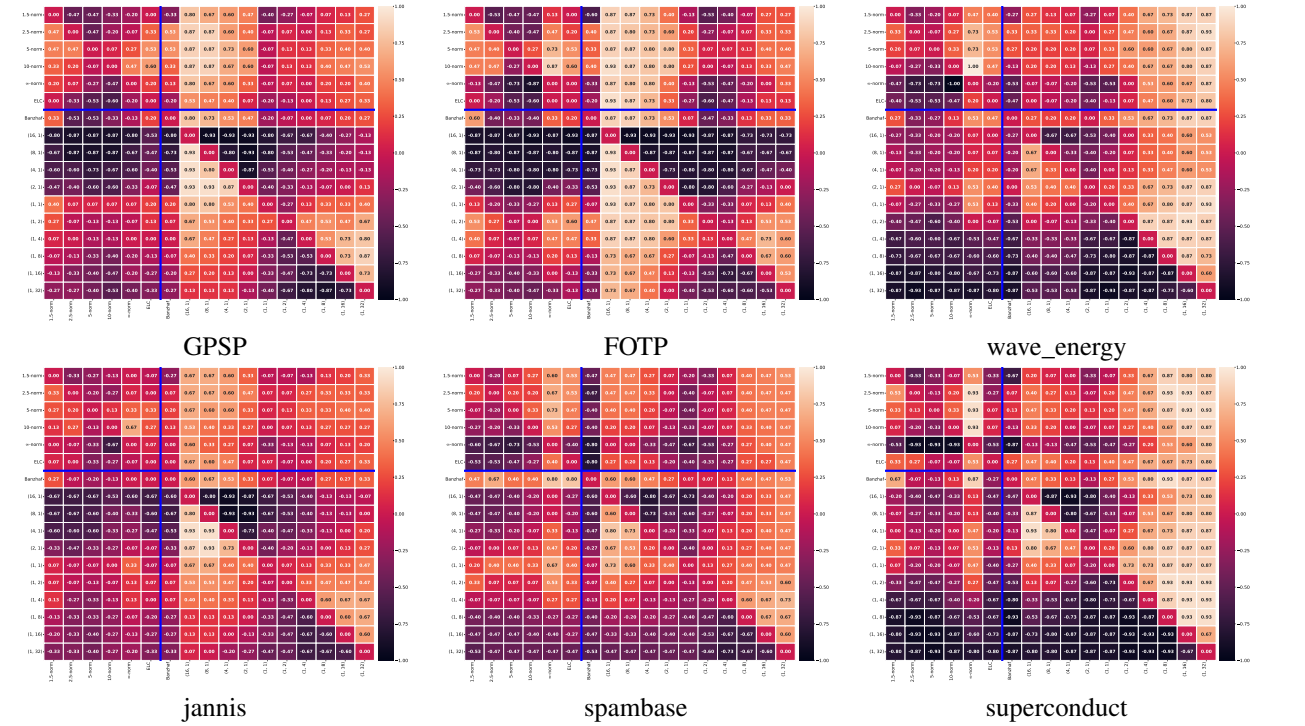

Figure 5: Statistical comparison of attribution methods on six datasets for player ranking. Since all utility functions contain more than 30 players, nonlinear methods are approximated by sampling $\#\text{players} \times 1,000$ subsets, with AUC averaged over five random seeds. Linear methods are computed exactly. The blue lines separate the linear and nonlinear methods.

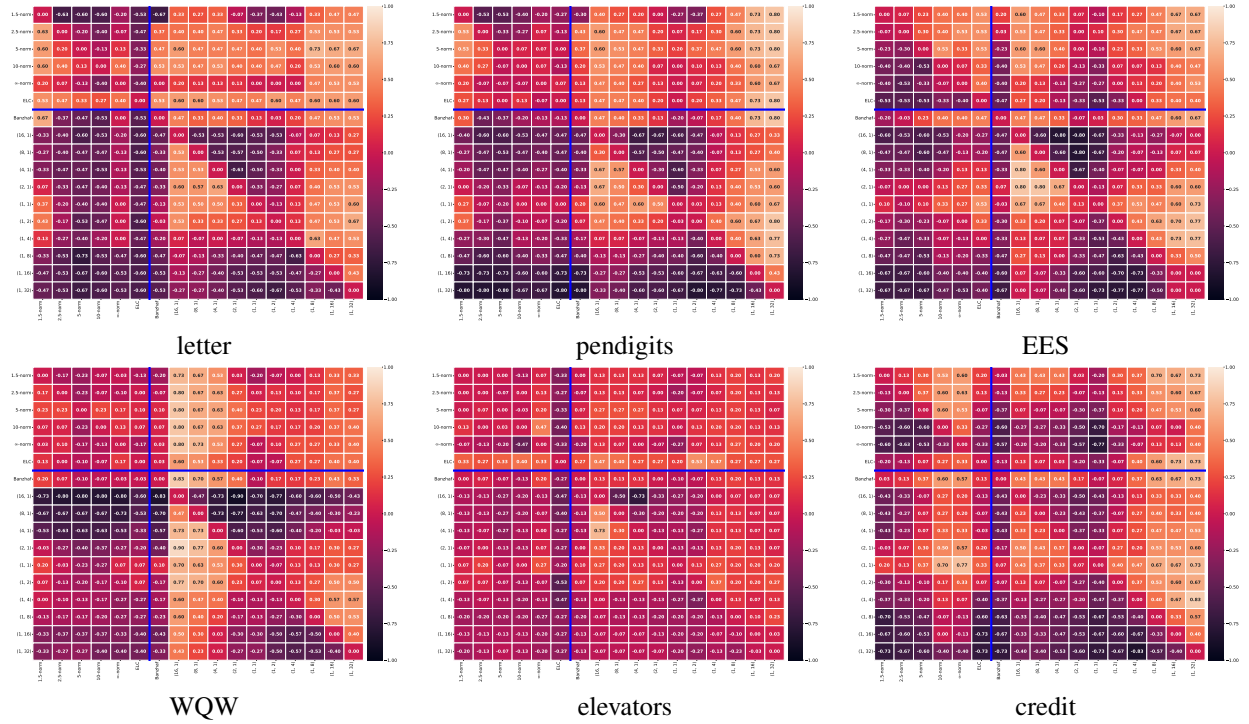

Figure 6: Statistical comparison of attribution methods on six datasets for player ranking. Since all utility functions have fewer than 20 players, all subsets are enumerated exactly for all methods. The blue lines separate the linear and nonlinear methods.