Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Survey Track

Approximation Algorithms for the Shapley Value: Taxonomy and Properties

Patrick Kolpaczki[1,2] , **Eyke Hüllermeier**[1,2,3]
[1]LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]German Research Center for Artificial Intelligence (DFKI, DSA)

Abstract

Attributing importance to the individual components of a larger unit has become a popular method for understanding models and data in AI and machine learning. Starting with feature explanation, this method is now also used in data valuation or federated learning, just to name a few. Despite their differences, all of these applications use the same mathematical attribution mechanism: the Shapley value, which is rooted in cooperative game theory. While the Shapley value is appealing and has strong axiomatic foundations, it is computationally intractable due to the combinatorial explosion of player subsets. Therefore, there is a need for approximation algorithms, which have been studied intensively in recent years. This survey provides an overview of general-purpose approximation methods applicable to any domain. We categorize these methods into algorithmic classes, compare their properties, and highlight connections between approaches in a comprehensive taxonomy.

1 Introduction

The notion of a cooperative game is omnipresent and appears in a wide range of fields far beyond its origin of economics. Multiple players solve together a task through cooperation and receive a collective reward that is to be shared among them. Not only would a fair distribution of the reward describe each player's contribution and reward it appropriately, the presence of such a mechanism itself can arguably incentivize collaboration in the first place. The field of machine learning adopted this paradigm, initially to attribute importance to individual features which contribute as players to the generalization performance of a learner [Covert *et al.*, 2020] or to the prediction for a specific instance [Lundberg and Lee, 2017], which are then divided among the feature to obtain importance scores. These scores form an explanation of how the features impact a model's behavior. In light of the rapid increase in complexity of machine learning models, eventually reaching opaqueness and black box nature, combined with the recent demand for trustworthiness of employed models, especially the research branch of explainable AI picked up this methodology to generate *feature explanations* that seek to provide understanding to the human user.

Simultaneously, this development made a frequently used payoff mechanism for deriving importance scores prominent in machine learning: the well-known *Shapley value* [Shapley, 1953]. It assigns to each player a payoff based on its average *marginal contribution*, expressing how much the inclusion of that player to a coalition of other players increases collective benefit. Compared to other alternatives, it uniquely satisfies axioms that arguably capture a widespread understanding of fairness, reinforcing its popularity.

Convincing with this appeal, the Shapley value has found various applications in machine learning that go beyond the well-studied purpose of generating *local* or *global* feature explanations [Sundararajan and Najmi, 2020; Covert *et al.*, 2021; Molnar, 2022]. Given the importance scores provided by the Shapley value, it comes naturally to conduct *feature selection* by favoring those with higher scores [Cohen *et al.*, 2007; Sebastián and González-Guillén, 2024]. However, this approach to perform selection is not limited to features. For instance, one can assign scores to data points in a training set to quantify the utility of each single data point. On this basis, the Shapley value is used for *data valuation* [Ghorbani and Zou, 2019; Jia *et al.*, 2019; Wu *et al.*, 2023], filtering out harmful outliers and reducing the training set size in general. In a similar vein, one can attribute in *federated learning* how much predictive performance data owners provide to a joint model [Sim *et al.*, 2020; Liu *et al.*, 2022]. With the subtle difference being that not individual data points are seen as players but whole datasets, each provided by a data owner. In contrast to explanations, the Shapley value provides here an economically-flavored purpose. It rewards clients appropriately for sharing their data, eventually incentivizing participation in the joint learning process. Lastly, one can select not only entities in the data but also within the model itself, thus arriving at *model pruning*. Akin to data valuation, one can remove base learners with negative scores from ensembles or in general select a significantly smaller subset [Rozemberczki and Sarkar, 2021]. Likewise, the Shapley value is used to prune artificial neural networks by removing single neurons depending on their scores [Ghorbani and Zou, 2020]. Moreover, it is also used for its explanatory character, providing an insight on how the neurons influence prediction. We refer to [Rozemberczki *et*

al., 2022] for an overview of use cases in machine learning.

Independent of the actual domain of interest, the Shapley value comes with a considerable drawback, impeding its application. Since a player's Shapley value takes into account its marginal contributions to all feasible subsets of players, its exact computation becomes quickly intractable, severely limiting the potential number of players (and thus features, data points, etc.) that can be considered. In particular, this exponential blowup provably results in NP-hardness for specific cooperative games [Deng and Papadimitriou, 1994]. As a viable alternative to exact computation, a diverse landscape of approximation algorithms has been developed. This subfield has attracted considerable interest in recent years as the importance of reliable and quick estimates for the Shapley value has been recognized, facilitating precise scores to quantify feature importance or pay out data owners. In contrast, closed-form polynomial solutions of the Shapley value have been found only for a small number of model classes in combination with specific formulations of collective benefit, such as decision trees [Lundberg *et al.*, 2020] or graph neural networks [Muschalik *et al.*, 2025].

Contribution. As the literature on approximating the Shapley value has grown in parallel to its spreading relevance in machine learning, we seek to provide an overview of the recently developed methodological landscape and thus solidify the progress made. Categorizing algorithms by their approach and utilized representation of the Shapley value, as well as pointing out connections between them, should provide a better understanding of the field and hopefully spark perspectives for novel work. Although surveys in that direction exist, we intend to contribute in a more conceptual, and universally applicable style that reaches a broader audience not limited to narrow machine learning tasks. Existing surveys are tailored towards feature explanations [Chen *et al.*, 2023; Olsen *et al.*, 2024], put more weight on model-specific estimation methods [Chen *et al.*, 2023], or emphasize a certain algorithmic class without providing a general taxonomy [Lamothe and Ngueveu, 2025]. We distinguish ourselves from these by considering methods that are domain-independent and can handle any cooperative game independent of how collective benefit is modeled. This implies the absence of structural assumptions or leveraging domain-specific heuristics. We believe that this approach appropriately reflects the wide span of applications for the Shapley value.

2 Cooperative Games and the Shapley Value

A cooperative game $(\mathcal{N}, \nu)$ consists of a *player set* $\mathcal{N}$ and a *value function* $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$. The player set contains the n involved players in the game and hence, we write $\mathcal{N} = \{1, \dots, n\}$ to keep that abstract nature. These can represent arbitrary objects of interest such as features, data points, or even base learners. To reflect that players can collaborate by forming groups to solve a task, subsets $S \subseteq \mathcal{N}$ are called *coalitions*. The collective benefit that each coalition could generate is specified by the value function as its *worth* $\nu(S)$. Thus, ν is the central component to model a cooperative game. Assuming that all players are willing to cooperate and form the *grand coalition* $\mathcal{N}$, one is confronted with the question of how to fairly distribute the generated worth $\nu(\mathcal{N}) - \nu(\emptyset)$ among them. The Shapley value [Shapley, 1953] is arguably the most well-known solution and assigns to each player $i \in \mathcal{N}$ the payoff

$$\phi_i := \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot \left[\nu(S \cup \{i\}) - \nu(S)\right]. \quad (1)$$

It can be interpreted as the weighted average of a player's *marginal contributions* $\Delta_i := \nu(S \cup \{i\}) - \nu(S)$ to all coalitions $S \subseteq \mathcal{N}^{-i} := \mathcal{N} \setminus \{i\}$ to which i could possibly contribute. If it is unclear for which value function ν the Shapley value is considered, we write ϕ_i^ν, otherwise in case of no ambiguity we use ϕ_i. The Shapley value's popularity is rooted in its axiomatic uniqueness as it is the only solution concept to fulfill four desiderata [Shapley, 1953] that are commonly understood to capture fairness:

- **Symmetry:** Two players with equal marginal contributions, i.e. $\Delta_i(S) = \Delta_j(S)$ for all $S \subseteq \mathcal{N} \setminus \{i, j\}$, receive the same payoff $\phi_i = \phi_j$.
- **Dummy player:** A player with no contribution, i.e. $\Delta_i(S) = 0$ for all $S \subseteq \mathcal{N}^{-i}$, receives no payoff: $\phi_i = 0$.
- **Linearity:** Scaling ν by a constant $c \in \mathbb{R}$ scales each player's payoff by c, i.e. $\phi_i^{\nu'} = \phi_i^\nu \cdot c$ for the cooperative game $(\mathcal{N}, \nu' := \nu \cdot c)$ for all $i \in \mathcal{N}$. Each player's payoff in the game $(\mathcal{N}, \nu_1 + \nu_2)$ resulting from the addition of two games $(\mathcal{N}, \nu_1)$ and $(\mathcal{N}, \nu_2)$, is the sum of the player's payoffs in both games, i.e. $\phi_i^{\nu_1 + \nu_2} = \phi_i^{\nu_1} + \phi_i^{\nu_2}$.
- **Efficiency:** The sum of all player's payoffs equals the caused collective benefit, i.e. $\sum_{i \in \mathcal{N}} \phi_i = \nu(\mathcal{N}) - \nu(\emptyset)$.

3 Approximating Shapley Values

The price paid for its convenient and unique properties is the computational intractability of the Shapley value. Its definition already reveals that in general all 2^n many coalitions need to be evaluated for their worth to compute ϕ_i exactly. Moreover, its exact computation is provably NP-hard for certain structured games whose value function can be specified in polynomial space w.r.t. n [Deng and Papadimitriou, 1994]. This exponential blow-up calls for a circumvention if one seeks to utilize the Shapley value in practice, especially in machine learning with dozens of features and hundreds if not thousands of data points that act as players. One typically addresses this problem by reliably estimating ϕ_i while keeping the number of evaluations low. In the following, we clarify our considered setting for approximation and list properties by which approximation algorithms are to be distinguished.

3.1 Approximation Problem

We assume a cooperative game $(\mathcal{N}, \nu)$ with unknown Shapley values $\phi := (\phi_1, \dots, \phi_n)$ to be given such that $\nu(S)$ can be evaluated for each coalition S by an approximation algorithm $\mathcal{A}$. The algorithm is expected to output estimates $\hat{\phi} := (\hat{\phi}_1, \dots, \hat{\phi}_n)$ with each $\hat{\phi}_i$ being close to ϕ_i, i.e. $\hat{\phi}_i \approx \phi_i$. One could also be interested in the estimation of a single particular ϕ_i, however, we stick to the approximation

Algorithm	Parameterless	Space complexity	Consistent	Unbiased	Expected MSE
[Castro *et al.*, 2009]	Yes	$\mathcal{O}(n)$	Yes	Yes	$\mathcal{O}(n/T)$
[Maleki *et al.*, 2013]	Yes	$\mathcal{O}(n^2)$	Yes	Yes	$\mathcal{O}(n/T)$
[van Campen *et al.*, 2018]	Yes	$\mathcal{O}(n^2)$	Yes	Yes	$\mathcal{O}(n/T)$
[Castro *et al.*, 2017]	No	$\mathcal{O}(n^2)$	Yes	Yes	–
[O'Brien *et al.*, 2015]	No	$\mathcal{O}(n^2)$	Yes	Yes	–
[Burgess and Chapman, 2021]	No	$\mathcal{O}(n^2)$	Yes	Yes	–
[Okhrati and Lipani, 2020]	No	$\mathcal{O}(nm)^a$	Yes / No^b	No	–
[Zhang *et al.*, 2023]	Yes	$\mathcal{O}(n)$	Yes	Yes	$\mathcal{O}(n/T)$
[Covert *et al.*, 2019]	Yes	$\mathcal{O}(n)$	Yes	Yes	✓
[Fumagalli *et al.*, 2023]	Yes	$\mathcal{O}(n)$	Yes	Yes	✓
[Kolpaczki *et al.*, 2024a]	Yes	$\mathcal{O}(n^2)$	Yes	Yes	$\mathcal{O}(\log n/T)$
[Kolpaczki *et al.*, 2024b]	No	$\mathcal{O}(n^2)$	Yes	Yes	–
[Lundberg and Lee, 2017]	Yes	$\mathcal{O}(n+T)$	Yes	–	–
[Covert and Lee, 2021]	No	$\mathcal{O}(n+T)$	Yes	Yes	✓^{c}
[Simon and Vincent, 2020]	No	$\mathcal{O}(n+T)$	–	–	✓
[Musco and Witter, 2025]	Yes	$\mathcal{O}(n+T)$	Yes	–	✓
[Chen *et al.*, 2025]	Yes	$\mathcal{O}(n+T)$	Yes	–	–
[Yan *et al.*, 2020]	No	$\mathcal{O}(n^k+T)$	Yes	–	✓
[Pelegrina *et al.*, 2025]	No	$O(n^k+T)$	Yes	–	–
[Witter *et al.*, 2025]	No	$\mathcal{O}(n+T)^d$	Yes	Yes	✓
[Corder and Decker, 2019]	No	$\mathcal{O}(n)$	Yes	No	–

Table 1: Properties of selected methods. Methods that sample without replacement and compute ϕ_i exactly for $T = 2^n$ are treated as consistent. Expected MSE is only taken w.r.t. n and T and no game-specific constants. '–' indicates no known result, '✓' indicates existing or trivial result but exact dependency on n and T yet unknown. [a] With m being the chosen number of equidistant points in $[0, 1]$. [b] Consistent if m grows with n without converging. [c] Due to equivalence with *SHAP-IQ*. [d] If chosen regression and Monte Carlo estimator have $O(n+T)$ space complexity.

of all player's Shapley values as this is commonly the case in machine learning. $\mathcal{A}$ has a *budget* of T many evaluations at its disposal, limiting how much of ν it can observe. Thus it selects a potentially random *retrieval sequence* $(S_1, \ldots, S_T)$ of coalitions to be evaluated and used to compute $\hat{\phi}_i$. The budget captures that access to ν typically poses the bottleneck in runtime, often requiring costly model inference or retraining in machine learning.

The quality of an estimate $\hat{\phi}_i$ is quantified by an error measure to be minimized, most commonly the mean squared error (MSE). But since it is usually of stochastic nature due to random sampling performed by $\mathcal{A}$, one usually judges $\mathcal{A}$ by its *expected MSE*:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right]. \tag{2}$$

Alternatives to the expected MSE exist by taking the absolute error and orthogonally to that choice divide it by the absolute Shapley value to obtain a proportional error [van Campen *et al.*, 2018]. Regarding the handling of $\mathcal{A}$'s randomness, one can instead of taking the expected error, consider for each player i the *error probability* by which $\mathcal{A}$ outputs an estimate that deviates by more than some $\varepsilon \geq 0$, usually expressed by an upper bound $1 - \delta$ with $\delta \in [0, 1]$ to be maximized:

$$\mathbb{P}(|\hat{\phi}_i - \phi_i| > \varepsilon) \leq 1 - \delta\,. \tag{3}$$

3.2 Properties

Besides the quality of the returned estimates, one can describe further characteristics and distinguish algorithms by other criteria, which are relevant before or during execution.

- **Parameter specification:** Some algorithms require the specification of additional parameters that impact the random generation of the retrieval sequence or the computation of $\hat{\phi}$. Often, the ideal parameterization depends on the considered game.
- **Space complexity:** Despite returning only n many point estimates $\hat{\phi}_i$, some algorithms do not exhibit a space complexity of $\mathcal{O}(n)$, but rather $\Omega(n^2)$. In general, at most a polynomial growth w.r.t. n is required for practicality. Some approaches even show a dependence of T.
- **Adaptivity:** The steps executed by $\mathcal{A}$ can be determined at initialization or during the approximation procedure by adapting to observations made by palpating ν. We term the former type of algorithms *static* and the latter *adaptive*. Typically, adaptive algorithms alter their sampling distribution, while their static counterparts have a priori fixed retrieval sequences.
- **Consistency:** For each $i \in \mathcal{N}$, the budget-dependent estimate $\hat{\phi}_i^T$ converges in probability towards ϕ_i, i.e. $\lim_{T\to\infty}\mathbb{P}(|\hat{\phi}_i^T - \phi_i| > \varepsilon) = 0$ for all $\varepsilon > 0$. Loosely speaking, consistency promises smaller estimation error with increasing budget.

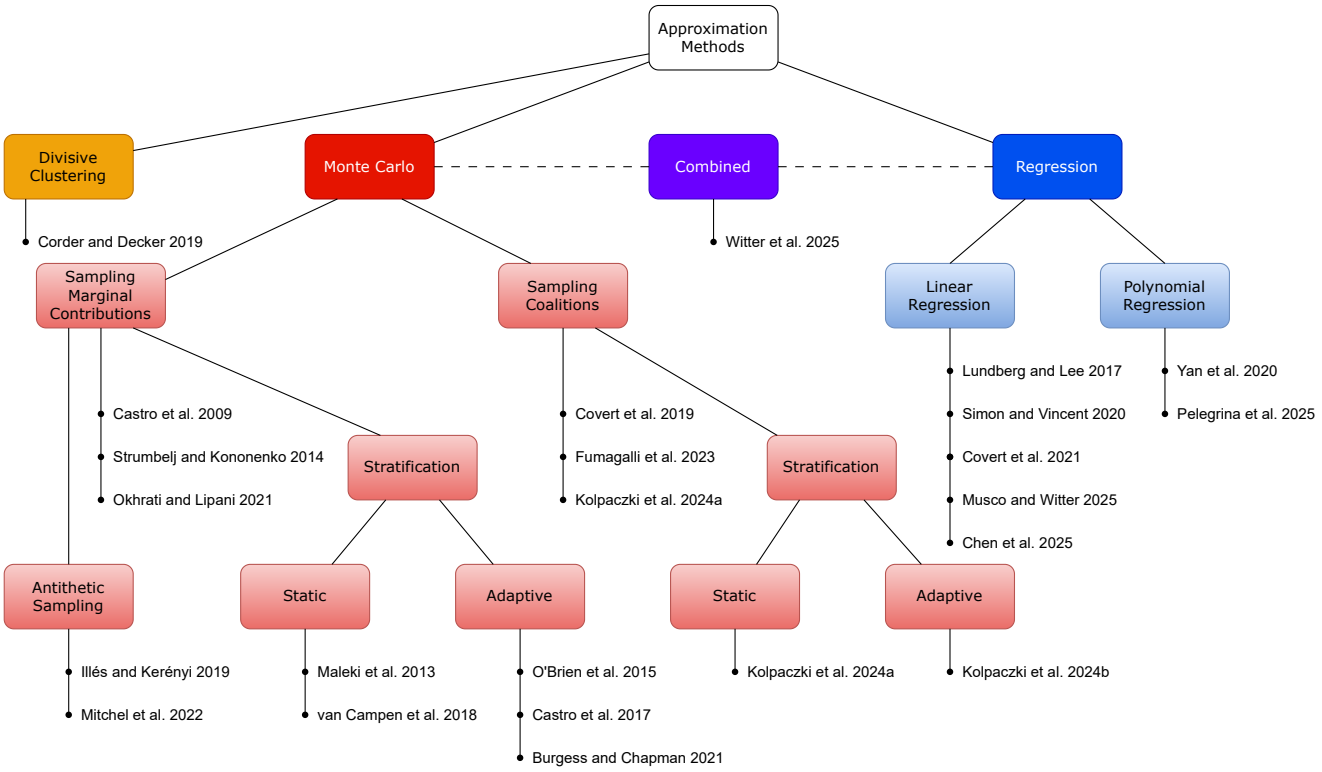

Figure 1: Taxonomy of selected domain-agnostic approximation methods for Shapley values grouped by algorithmic classes.

- **Unbiasedness:** For each $i \in \mathcal{N}$, the estimates equal their Shapley value in expectation, i.e. $\mathbb{E}[\hat{\phi}_i] = \phi_i$. Particularly for explanations, unbiased estimates are desirable to avoid inherently misleading interpretations. Notably, biased estimates can still be consistent when the bias shrinks with T.
- **Error bounds:** Theoretical results regarding the estimation error have been shown for a variety of algorithms. Usually these depend on n and T as well as latent constants of the game, such as the variance of the population from which $\mathcal{A}$ samples. Their benefit is two-fold: on the one hand they hint at which algorithms are preferable for certain types of games, on the other, understanding how the shape of ν in combination with $\mathcal{A}'s$ sampling procedure impacts estimation error can guide algorithm design to further improvements.

Table 1 lists algorithms and their properties. In the following sections, we present the developments within algorithmic classes, leading to a taxonomy illustrated in Figure 1, which portrays the advance of this research field.

4 Monte Carlo Methods

The arguably most widely studied class of approximation algorithms is based on a likely evident observation: the Shapley value can be interpreted as a player's expected marginal contribution, i.e. $\phi_i = \mathbb{E}[\Delta_i(S)]$ w.r.t. a probability distribution over $\mathcal{P}(\mathcal{N}^{-i})$ given by the weights in (1). Hence, it comes natural to reduce Shapley approximation to mean estimation and apply Monte Carlo methods. This class appeals with many merits: its members are mostly (trivially) consistent, unbiased, and have often error bounds for expected MSE and error probability.

4.1 Sampling Marginal Contributions

The first representative, *ApproShapley* [Castro *et al.*, 2009], samples marginal contributions for each player which it elicits from permutations of $\mathcal{N}$. Within a permutation π for each player i the marginal contribution $\Delta_i(\text{pre}_i(\pi))$ can be taken to update $\hat{\phi}_i$ where $\text{pre}_i(\pi)$ is the coalition of players preceeding i in π. This method is unbiased as it directly approximates a representation of π based on permutations:

$$\phi_i = \sum_{\pi \in \Pi} \Delta_i(\text{pre}_i(\pi)) \tag{4}$$

with Π being the set of all permutations over $\mathcal{N}$. This procedure is budget efficient since the evaluated worth $\nu(\text{pre}_i(\pi) \cup \{i\})$ can be reused for the next player's marginal contribution. However sampling marginal contributions separately for each player allows for varying sample numbers m_i among players. Thus, one can prioritize players with higher sample variance $\hat{\sigma}_i^2$. More specifically, selecting the next player greedily, by sample the one who promises maximal estimated MSE reduction $\hat{\sigma}_i^2/m_i - \hat{\sigma}_i^2/m_i+1$ [Strumbelj and Kononenko, 2014], constitutes the first adaptive algorithm we encounter.

Given that $\mathcal{A}$ is unbiased, the expected MSE boils down to $1/n \sum_{i \in \mathcal{N}} \mathbb{V}[\hat{\phi}_i]$. Hence, applying variance reduction tech-

niques is a natural path of improvement. The most commonly applied technique is that of *stratification*. Here, one partitions a population into *strata* and samples from each strata independently. The obtained strata estimates are then aggregated to an estimate of the whole population. The underlying idea is that the strata estimates should converge quickly if each strata showcases a high degree of homogeneity within its subpopulation, thus leading to a final estimate of lower variance. For each ϕ_i, stratifying is carried out by forming a stratum for each coalition size s to which i has a marginal contribution, resulting in n strata per player and n^2 in total [Maleki *et al.*, 2013]. The Shapley value even admits a representation as an arithmetic mean of strata values $\phi_{i,s}$, each of which is itself an arithmetic mean of marginal contributions:

$$\phi_i = \frac{1}{n} \sum_{s=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{s}} \sum_{S \subseteq \mathcal{N}^{-i}, |S|=s} \Delta_i(S)}_{:=\phi_{i,s}} . \tag{5}$$

A crucial role of this approach is played by the *sample allocation* over all strata specified by the number of samples $m_{i,s}$ drawn for each $\phi_{i,s}$. In its simplest form, the sample allocation is uniform as emulated by *Structured Sampling* [van Campen *et al.*, 2018]. A randomly drawn permutation of $\mathcal{N}$ is duplicated multiple times and altered, such that each player appears exactly once at each position, from which then marginal contributions are retrieved as in [Castro *et al.*, 2009]. More advanced methods perform *adaptive stratification*. Based on the observation that stratification yields an expected MSE[1] of

$$\frac{1}{n^3} \sum_{i=1}^{n} \sum_{s=0}^{n-1} \frac{\sigma_{i,s}^2}{m_{i,s}} \tag{6}$$

with stratum variances $\sigma_{i,s}^2$, one can derive the optimal allocation that minimizes MSE:

$$m_{i,s}^* = \frac{\sigma_{i,s} \cdot T}{2 \sum_{j=1}^{n} \sum_{k=1}^{n-1} \sigma_{j,k}} , \tag{7}$$

assuming that each sample costs two evaluations, which is known as the *Neyman allocation* [Neyman, 1934]. Since the stratum variances are unknown, adaptive methods estimate them and adjust their sample allocation accordingly during approximation with the goal of converging to $m_{i,s}^*$. Pursuing this idea, adaptive methods face an exploration-exploitation dilemma in which exploration expresses gathering information about $\sigma_{i,s}^2$ through sampling all strata and exploitation consists of chasing the estimated optimal allocation. The *two-phase approach* implements a simple "first-explore-then-exploit" strategy by first sampling uniformly from all strata and then trying to reach the estimated optimal allocation in the second phase [Castro *et al.*, 2017]. In contrast, employing a sigmoid function to randomly pick between exploration and exploitation, slowly converging to the latter, represents a smoother transition between the two phases [O'Brien *et al.*, 2015]. Both require parameterization, be it the phase lengths or the sigmoid function. The *Stratified Empirical Bernstein Method* greedily selects the next stratum to sample from by minimizing an upper bound on the error probability [Burgess and Chapman, 2021]. However, it requires the knowledge of the ranges of values within the strata. While these methods are unbiased, analytical results on their achieved error remain unknown. A potentially significant drawback is the number of strata scaling quadratically with n. Besides increased space complexity this poses a vulnerability for adaptive stratification. The accuracy of the estimated optimal sample allocation over the strata suffers if the available budget is not sufficient to estimate startum avriance precisely, and thus the algorithm is mislead during exploitation.

Another variance reduction technique is that of *antithetic sampling*. The sampled marginal contributions for each player are not drawn independent anymore but constructed in such a way that these exhibit (likely) negative covariance [Illés and Kerényi, 2019; Mitchell *et al.*, 2022].

Interestingly, the Shapley value can not only be expressed as a discrete sum, but also as an integral called *multilinear extension* [Owen, 1972]:

$$\phi_i = \int_0^1 \mathbb{E}[\Delta_i(S_{i,q})] \, dq \, , \tag{8}$$

with distribution $\mathbb{P}(S_{i,q} = S) = q^{|S|} \cdot (1-q)^{n-|S|-1}$ over $\mathcal{P}(\mathcal{N}^{-i})$. Thus, one can approximate ϕ_i through numerical integration, for example by estimating $\mathbb{E}[\Delta_i(S_{i,q})]$ for equidistant points $q \in [0,1]$ [Okhrati and Lipani, 2020]. This introduces a trade-off specified by parameterization between the number of points and the number of samples left to obtain precise estimates. Further, the equidistant points introduce a bias. Worth mentioning is the connection to stratification: instead of estimating $\mathbb{E}[\Delta_i(S_{i,q})]$ for random $q \in [0,1]$, the unit interval is partitioned with each q representing a stratum.

Slightly deviating from marginal contributions, one can switch the sampled unit to complementary contributions $\Delta_i^C(S) := \nu(S \cup \{i\}) - \nu(\mathcal{N} \setminus (S \cup \{i\}))$. These can likewise be averaged to obtain

$$\phi_i = \sum_{S \subseteq \mathcal{N}^{-i}} \frac{1}{n\binom{n-1}{|S|}} \cdot \Delta_i^C(S) \, , \tag{9}$$

allowing one to analogously transfer previous approaches onto this representation [Zhang *et al.*, 2023].

4.2 Sampling Coalitions

Despite the fundamental role of the marginal contribution, the Shapley value can be represented through other building blocks giving rise to methods that sample these units instead. Single coalitions constitute the most atomic building block allowing improved sampling efficiency. One can rewrite ϕ_i as

$$\phi_i = \sum_{S \subseteq \mathcal{N}} \omega_{i,S} \cdot \nu(S) \tag{10}$$

with $\omega_i = \frac{1}{n\binom{n-1}{|S|-1}}$ for $i \in S$ and $\omega_i = -\frac{1}{n\binom{n-1}{|S|}}$ otherwise. *SHAP-IQ* [Fumagalli *et al.*, 2023] samples coalitions

[1] For simplicity we neglect that the most outer strata $s = 0$ and $s = n-1$ contain only one element and thus have zero variance.

from $\mathcal{P}(N)$, evaluates them, and updates all $\hat{\phi}_i$ via *importance sampling*. Noteworthy is the usage of each evaluated worth $\nu(S)$ for all players' estimates because it is part of each player's representation of ϕ_i. This *maximum sample reuse* principle [Wang and Jia, 2023] is unattainable for approaches based on marginal contributions since each observed $\Delta_i(S)$ belongs only to i itself and no other player. Leveraging this advantage, further methods have been proposed, for example the disentanglement of coalition values with differing signs into two sums resulting in:

$$\phi_i = \underbrace{\sum_{S \subseteq \mathcal{N}^{-i}} \frac{1}{n\binom{n-1}{|S|}} \nu(S \cup \{i\})}_{=:\phi_i^+} - \underbrace{\sum_{S \subseteq \mathcal{N}^{-i}} \frac{1}{n\binom{n-1}{|S|}} \nu(S)}_{=:\phi_i^-} . \quad (11)$$

Instead of estimating each average separately [Kolpaczki *et al.*, 2024a] making error analyses accessible, one can perform importance sampling from one distribution over $\mathcal{P}(\mathcal{N})$ and use each evaluated $\nu(S)$ to update either $\hat{\phi}_i^+$ or $\hat{\phi}_i^-$ for each player i [Covert *et al.*, 2019]. Provably, there exists no single distribution over $\mathcal{P}(\mathcal{N})$ that leads to unbiased estimates without importance sampling [Kolpaczki *et al.*, 2024a]. The separation into ϕ_i^+ and ϕ_i^- can be understood as an attempt to stratify, however, one can pursue this direction even further by additionally considering coalition size:

$$\phi_i = \frac{1}{n}\sum_{s=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{s}} \sum_{S \subseteq \mathcal{N}^{-i}, |S|=s} \nu(S \cup \{i\})}_{=:\phi_{i,s}^+} - \frac{1}{n}\sum_{s=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{s}} \sum_{S \subseteq \mathcal{N}^{-i}, |S|=s} \nu(S)}_{=:\phi_{i,s}^-} . \quad (12)$$

The aggregation of stratum values to ϕ_i can be simplified to

$$\phi_i = \frac{1}{n}\sum_{s=0}^{n-1} \phi_{i,s}^+ - \phi_{i,s}^- . \quad (13)$$

The applied stratification does not only serve as a variance reduction technique by providing the benefit of more homogeneous strata. It further allows one to recover the maximum sample reuse principle without the need of importance sampling and thus facilitating error analyses. In *Stratified SVARM* [Kolpaczki *et al.*, 2024a], this is achieved by observing that each stratum value is an arithmetic mean of coalition values. Hence, each coalition sampled from a distribution over $\mathcal{P}(\mathcal{N})$ is used for every player to update an estimate of the stratum to which it belongs. The stratification can be further extended to adaptivity by the techniques presented in Section 4.1. For example, *Adaptive SVARM* [Kolpaczki *et al.*, 2024b] implements the two-phase approach [Castro *et al.*, 2017].

5 Regression Methods

A different class of algorithms obtains estimates by solving an optimization problem, essentially performing regression.

5.1 Linear Regression

The groundlaying result for this stream of works is the discovery that the Shapley value itself is a solution to a constrained weighted least squares (WLS) problem [Charnes *et al.*, 1988]:

$$\phi_i = \underset{\phi' \in \mathbb{R}^n, \phi_0' \in \mathbb{R}}{\arg\min} \sum_{S \in \mathcal{P}(\mathcal{N}) \setminus \{\emptyset, \mathcal{N}\}} \frac{1}{\binom{n-2}{|S|-1}} \left(\nu(S) - \phi_0' - \sum_{i \in S} \phi_i' \right)^2$$
$$\text{s.t.} \sum_{i=1}^{n} \phi_i' = \nu(\mathcal{N}) - \nu(\emptyset) \quad (14)$$

At first, this formulation leads back to an exponential sum in the objective function. However, one can circumvent this hurdle by gradually constructing it from observed coalitions.

Kernel weights. *KernelSHAP* [Lundberg and Lee, 2017] samples and evaluates coalitions with replacement to add each observed pair S and $\nu(S)$ to the sum in the objective function. The solution to the smaller WLS problem is then treated as the desired estimate $\hat{\phi}$. This approach comes at the cost of space complexity depending on T since each evaluated $\nu(S)$ must be stored. Moreover, the analysis of error and other properties is challenging [Covert and Lee, 2021]. In contrast, *Unbiased KernelSHAP* [Covert and Lee, 2021] alters the solution to the WLS problem on an algebraic level and is provably unbiased. Surprisingly, the resulting estimates are equivalent to those of *SHAP-IQ* [Fumagalli *et al.*, 2023], implying a connection between both major algorithmic classes. Further, the technique of *paired sampling* is proposed, additionally evaluating the complement S^C of each sampled S. Given the convexity of the optimization problem, methods that still sample but conduct stochastic gradient descent have been proposed [Simon and Vincent, 2020].

Leverage scores. Less considered, but still an element of vital importance, is the distribution from which coalitions are sampled to construct the estimated objective function. *KernelSHAP* heuristically samples with probability proportionally to the weights $\binom{n-2}{|S|-1}$ in (14) known as *kernel weight sampling*. Instead, *Leverage SHAP* [Musco and Witter, 2025] performs *leverage sampling* based on leverage scores originally quantifying the impact of individual matrix rows in regression problems. Considering (14) on an algebraic level where each coalition responds to a row of a matrix allows one to transfer this paradigm. This further unveils a connection to active learning in which leverage scores are used for the selection of the next data point. Likewise, estimation via regression can be seen as an active learning problem in which the coalitions must be chosen. As a result, each coalition of size s is drawn with probability proportional to $\binom{n}{s}^{-1}$ derived from a rigorous analysis [Musco and Witter, 2025] that reduces the complexity of computing scores naively in time $O(2^n n^2)$ to constant time. The bound on expected MSE of *Leverage SHAP* marks a breakthrough result for regression methods although the game-specific constants and asymptotic impact of n and T are relatively hard to interpret.

Modified row-norms. Both methods can be unified under *sketched regression* [Chen *et al.*, 2025] yielding a representation as unconstrained regression. Within this framework the first bound for *KernelSHAP* on the sample complexity guarantee an error probability is derived. Further, the geometric

mean of kernel weights and leverage scores as a sample distribution named *modified row-norms* [Chen *et al.*, 2025] shows favorable sample complexity bounds.

5.2 Generalized Regression

The WLS problem can be viewed from another perspective providing an interpretation that exhibits a more encompassing and unifying character. The term $\phi'_0 + \sum_{i \in S} \phi'_i$ can be seen as an estimate $\nu'(S)$ of $\nu(S)$. Hence, the regression fits a *surrogate game* $(\mathcal{N}, \nu')$ with $\nu'(S) = \phi'_0 + \sum_{i \in S} \phi'_i$ by minimizing the weighted mean squared error of ν' w.r.t. ν. Despite its lacking flexibility, evident in $\Delta_i(S) = \phi'_i$ for all $i \in \mathcal{N}$ and $S \subseteq \mathcal{N}^{-i}$, this surrogate game yields the exact solution when all coalitions are observed. Consequently, one may ask whether other classes of parameterizable surrogate games likewise yield the Shapley value and if increased flexibility obtained by a higher degree of freedom improves the fit during sampling.

Inspired by the notion of k-additivity, *SVA*k_{ADD} [Pelegrina *et al.*, 2025] provides a positive example by enriching the surrogate game with Shapley interactions [Grabisch and Roubens, 1999] up to degree k:

$$\nu'(S) = \sum_{S' \subseteq \mathcal{N}, |S'| \leq k} \gamma_{S,S'} \cdot \phi_{S'} \,, \tag{15}$$

where $\gamma_{S,S'}$ is a constant depending only on S and S'. Each interaction $\phi_{S'}$ captures the synergy arising from all players in S being present. While its correctness is proven for $k \leq 3$ and assumed for all $k \leq n$, space complexity and computational effort for solving obviously grow with k.

Another class of k-additive surrogate games are sums of unanimity games, effectively linear combinations of binary value functions, one for each coalition up to size k [Yan *et al.*, 2020].

6 Other Approaches

Interestingly, the two classes of Monte Carlo methods and regression methods are not necessarily in conflict with each other and metaphorically speaking can be combined to cooperate. This would simultaneously harness the unbiasedness of the former and the fast convergence of the latter (empirically in machine learning). The key observation behind this thought is the utility of the linearity axiom. A simple model $f \approx \nu$ or surrogate game in general with Shapley values computable in polynomial time is combined with a Monte Carlo method approximating the Shapley values of the *residual* $\nu - f$. Due to linearity, the addition of both estimates is expedient, since we have

$$\phi^{\nu} = \phi^{f} + \phi^{\nu - f} \,. \tag{16}$$

The universally valid decomposition of the value function into f and $\nu - f$ gives room to apply any model and mean estimation coined *RegressionMSR* [Witter *et al.*, 2025] and comes with theoretical guarantees: unbiasedness and a probabilistic bound on the MSE. Proposed candidates for f are linear models and XGBoost tree models [Witter *et al.*, 2025].

A vastly different approach is that of *divisive clustering* [Corder and Decker, 2019]. Instead of approximating the Shapley values directly for the considered cooperative game $(\mathcal{N}, \nu)$, the game is divided into smaller subgames for which the Shapley values are computed exactly, which is made feasible due to their smaller size. Interestingly, this division is not carried out on the level of ν but $\mathcal{N}$. For some parameter $\beta > 1$, the players are partitioned by the leaf nodes of a binary tree into mutually disjoint coalitions of size at most $\log_\beta(n)$. Each leaf node constitutes a subgame containing only the leaf node's players and ν projected onto this coalition. The exact Shapley values are computed for each subgame and interpolated bottom-up from the leaves to the root of the tree to ensure the efficiency axiom. For any parent node of two siblings S_1 and S_2 this is done through scaling the Shapley values of each subgame $S \in \{S_1, S_2\}$ by $\nu(S)\nu(S_1 \cup S_2) / \nu(S_1) + \nu(S_2)$ to maintain the proportion between S_1 and S_2. Sensitive parameters are the leaf node size and the partitioning mechanism.

7 Open Directions

The previous sections summarize the substantial progress made in the past decade on the topic of approximating the Shapley value. Nevertheless, the field still offers multiple opportunities for improvement and room for novel approaches.

Monte Carlo methods. We suspect that orthogonal variance reduction techniques can be further combined and applied onto different representations of the Shapley value. Moreover and specifically targeting stratification, the quadratic scaling of the space complexity w.r.t. n is a burden to be tackled. Caused by the sheer number of strata, one might introduce a coarser partitioning into less strata, eventually even depending on n and the available budget for adaptive methods. If one is interested in the problem-specific empirical impact, one could study how the game-specific constants that dictate the error of each technique and representation behave in varying domains such as feature explanations, data valuation, and model pruning, among others.

Regression methods. The unifying concept of fitting a surrogate game for regression methods lacks a deeper theoretical foundation. One may ask which parameterizable classes of games yield the Shapley value as a solution. Further, it is yet unclear how a better fit caused by a game's increased flexibility transfers into more precise estimates, or whether this is even necessarily the case. Regarding this class and its few theoretical results in general, error bounds with clearer dependencies on the number of players and budget and compact game-specific constants would certainly be of benefit. These would facilitate the comparison with approaches conducting mean estimation and give guidance on the selection of a suitable estimator for specific domains.

Alternative problem settings. A further research direction could be the tailoring of methods to more specific or slightly altered problem settings. Certain domains may come with patterns in the value function that are to be exploited by assumptions that a method is allowed to make. Variations may also be formulated regarding the player set. Eventually, selection tasks do not come with the rigid demand for precise estimates for all players. In this case, imprecision to some greater degree of less important players may be tolerated or some players may even completely drop out of consideration.

References

[Burgess and Chapman, 2021] Mark Alexander Burgess and Archie C. Chapman. Approximating the shapley value using stratified empirical bernstein sampling. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[Castro *et al.*, 2009] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 2009.

[Castro *et al.*, 2017] Javier Castro, Daniel Gómez, Elisenda Molina, and Juan Tejada. Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research*, 2017.

[Charnes *et al.*, 1988] Abraham Charnes, Boaz Golany, Michael Keane, and John Rousseau. *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*. 1988.

[Chen *et al.*, 2023] Hugh Chen, Ian Covert, Scott Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 2023.

[Chen *et al.*, 2025] Tyler Chen, Akshay Seshadri, Mattia J. Villani, Pradeep Niroula, Shouvanik Chakrabarti, Archan Ray, Pranav Deshpande, Romina Yalovetzky, Marco Pistoia, and Niraj Kumar. A unified framework for provably efficient algorithms to estimate shapley values. *CoRR*, abs/2506.05216, 2025.

[Cohen *et al.*, 2007] Shay B. Cohen, Gideon Dror, and Eytan Ruppin. Feature selection via coalitional game theory. *Neural Computation*, 2007.

[Corder and Decker, 2019] Kevin Corder and Keith Decker. Shapley value approximation with divisive clustering. In *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019.

[Covert and Lee, 2021] Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[Covert *et al.*, 2019] Ian Covert, Scott Lundberg, and Su-In Lee. Shapley feature utility. In *Machine Learning in Computational Biology*, 2019.

[Covert *et al.*, 2020] Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Covert *et al.*, 2021] Ian Covert, Scott M. Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 2021.

[Deng and Papadimitriou, 1994] Xiaotie Deng and Christos H. Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 1994.

[Fumagalli *et al.*, 2023] Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. Shap-iq: Unified approximation of any-order shapley interactions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[Ghorbani and Zou, 2019] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[Ghorbani and Zou, 2020] Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Grabisch and Roubens, 1999] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 1999.

[Illés and Kerényi, 2019] Ferenc Illés and Péter Kerényi. Estimation of the shapley value by ergodic sampling. *CoRR*, abs/1906.05224, 2019.

[Jia *et al.*, 2019] 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 *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[Kolpaczki *et al.*, 2024a] Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[Kolpaczki *et al.*, 2024b] Patrick Kolpaczki, Georg Haselbeck, and Eyke Hüllermeier. How much can stratification improve the approximation of shapley values? In *Proceedings of Explainable Artificial Intelligence - Second World Conference (xAI), Part II*, 2024.

[Lamothe and Ngueveu, 2025] François Lamothe and Sandra Ulrich Ngueveu. Approximating the Shapley value with sampling : survey and new stratification techniques. 2025.

[Liu *et al.*, 2022] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on Intelligent Systems and Technology*, 2022.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Lundberg *et al.*, 2020] Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex J. 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*, 2020.

[Maleki *et al.*, 2013] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *CoRR*, abs/1306.4265, 2013.

[Mitchell *et al.*, 2022] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 2022.

[Molnar, 2022] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[Muschalik *et al.*, 2025] Maximilian Muschalik, Fabian Fumagalli, Paolo Frazzetto, Janine Strotherm, Luca Hermes, Alessandro Sperduti, Eyke Hüllermeier, and Barbara Hammer. Exact computation of any-order shapley interactions for graph neural networks. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[Musco and Witter, 2025] Christopher Musco and R. Teal Witter. Provably accurate shapley value estimation via leverage score sampling. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[Neyman, 1934] Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 1934.

[O'Brien *et al.*, 2015] Gearóid O'Brien, Abbas El Gamal, and Ram Rajagopal. Shapley value estimation for compensation of participants in demand response programs. *IEEE Trans. Smart Grid*, 2015.

[Okhrati and Lipani, 2020] Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate shapley values. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, 2020.

[Olsen *et al.*, 2024] Lars Henry Berge Olsen, Ingrid Kristine Glad, Martin Jullum, and Kjersti Aas. A comparative study of methods for estimating model-agnostic shapley value explanations. *Data Min. Knowl. Discov.*, 2024.

[Owen, 1972] Guillermo Owen. Multilinear extensions of games. *Management Science*, 1972.

[Pelegrina *et al.*, 2025] Guilherme Dean Pelegrina, Patrick Kolpaczki, and Eyke Hüllermeier. Shapley value approximation based on k-additive games. *CoRR*, abs/2502.04763, 2025.

[Rozemberczki and Sarkar, 2021] Benedek Rozemberczki and Rik Sarkar. The shapley value of classifiers in ensemble games. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, 2021.

[Rozemberczki *et al.*, 2022] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Oliver Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[Sebastián and González-Guillén, 2024] Carlos Sebastián and Carlos E. González-Guillén. A feature selection method based on shapley values robust for concept shift in regression. *Neural Computing and Applications*, 2024.

[Shapley, 1953] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, 1953.

[Sim *et al.*, 2020] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[Simon and Vincent, 2020] Grah Simon and Thouvenot Vincent. A projected stochastic gradient algorithm for estimating shapley value applied in attribute importance. In *Proceedings of Machine Learning and Knowledge Extraction*, 2020.

[Strumbelj and Kononenko, 2014] Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 2014.

[Sundararajan and Najmi, 2020] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[van Campen *et al.*, 2018] Tjeerd van Campen, Herbert Hamers, Bart Husslage, and Roy Lindelauf. A new approximation method for the shapley value applied to the WTC 9/11 terrorist attack. *Social Network Analysis and Mining*, 2018.

[Wang and Jia, 2023] Jiachen T. Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[Witter *et al.*, 2025] R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted monte carlo estimators for shapley values and probabilistic values. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[Wu *et al.*, 2023] Mengmeng Wu, Ruoxi Jia, Changle Lin, Wei Huang, and Xiangyu Chang. Variance reduced shapley value estimation for trustworthy data valuation. *Computers & Operations Research*, 2023.

[Yan *et al.*, 2020] Tom Yan, Christian Kroer, and Alexander Peysakhovich. Evaluating and rewarding teamwork using cooperative game abstractions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Zhang *et al.*, 2023] Jiayao Zhang, Qiheng Sun, Jinfei Liu, Li Xiong, Jian Pei, and Kui Ren. Efficient sampling approaches to shapley value approximation. *Proc. ACM Manag. Data*, 2023.