Shapley Value Approximation Based on k-Additive Games

Guilherme Dean Pelegrina[*1], **Patrick Kolpaczki**[*2,3], **Eyke Hüllermeier** [2,3,4]
[1]Engineering School - Mackenzie Presbyterian University
[2]LMU Munich
[3]Munich Center for Machine Learning (MCML)
[4]German Research Center for Artificial Intelligence (DFKI, DSA)
guilherme.pelegrina@mackenzie.br, patrick.kolpaczki@lmu.de, eyke@lmu.de

Abstract

The Shapley value is the prevalent solution for fair division problems in which a payout is to be divided among multiple agents. By adopting a game-theoretic view, the idea of fair division and the Shapley value can also be used in machine learning to quantify the individual contribution of features or data points to the performance of a predictive model. Despite its popularity and axiomatic justification, the Shapley value suffers from a computational complexity that scales exponentially with the number of entities involved, and hence requires approximation methods for its reliable estimation. We propose *SVAk*$_{\text{ADD}}$, a novel approximation method that fits a k-additive surrogate game. By taking advantage of k-additivity, we are able to elicit the exact Shapley values of the surrogate game and then use these values as estimates for the original fair division problem. The efficacy of our method is evaluated empirically and compared to competing methods.

Code — https://github.com/GuilhermePelegrina/Shapley_Value_Approximation_Based_on_k-Additive_Games.git

Extended version — https://arxiv.org/abs/2502.04763

1 Introduction

The complexity of machine learning models experienced a rapid and certainly significant increase over the last decade. This development comes with an ever-rising burden to understand a model's decision-making, reaching a point beyond human comprehension. Meanwhile, societal and political influences led to a growing demand for trustworthy AI (Li et al. 2023). The field of Explainable AI (XAI) emerges to counteract these consequences, aiming to bring back understanding to the human user. Among various explanation types (Molnar 2021), post-hoc additive explanations convince with an intuitive appeal: an observed numerical effect caused by the behavior of the black box model is divided among participating entities. Additive feature explanations decompose a predicted value (Lundberg and Lee 2017) or generalization performance (Covert, Lundberg, and Lee 2020) among the involved features, enabling feature importance scores. Beyond explainability, this allows in feature engineering to conduct feature selection by removing features with irrelevant or even harmful contributions (Cohen, Ruppin, and Dror 2005; Marcílio and Eler 2020).

Treating this decomposition as a fair division problem opens the door to game theory which views the features as cooperating agents, forming groups called coalitions to achieve a task and collect a common reward that is to be shared. Such scenarios are captured by the widely applicable notion of cooperative games, modeling the agents as a set of players N and assuming that a real-valued worth $\nu(A)$ can be assigned to each coalition $A \subseteq N$ by a value function ν. Among multiple propositions the Shapley value (Shapley 1953) prevailed as the most favored solution to the fair division problem. It assigns to each player a share of the collective benefit, more precisely a weighted average of all its marginal contributions, i.e., the increase in collective benefit a player causes when joining a coalition. Its popularity is rooted in the fact that it is provably the only solution to fulfill certain desirable axioms (Shapley 1953) which arguably capture a widespread understanding of fairness. For example, in supply chain cooperation (Fiestras-Janeiro et al. 2011), the cost reduction when joining a coalition may be shared among companies based on the Shapley value. The greater a company's marginal contributions to the cost reduction, the greater its received payoff.

The applicability of the Shapley value exceeds economics as its utility has been recognized within various disciplines. Most prominently, it has found its way into the field of machine learning, especially as a model-agnostic approach, quantifying the importance of entities such as features or datapoints (see (Rozemberczki et al. 2022) for an overview). Adopting the game-theoretic view, these entities are understood as players which cause a certain numerical outcome. Shaping the measure of a coalition's worth adequately is pivotal to the informativeness of the importance scores obtained by the Shapley values. For example, considering a model's generalization performance on a test dataset restricted to the feature subset given by a coalition yields global feature importance scores (Pfannschmidt et al. 2016; Covert, Lundberg, and Lee 2020). Conversely, local feature attribution scores are obtained by splitting the model's prediction value for a fixed datapoint (Lundberg and Lee 2017). The Shapley value is not limited to provide additive explanations since it has also been proposed to perform data valuation (Ghorbani

[*]These authors contributed equally.

and Zou 2019), feature selection (Cohen, Dror, and Ruppin 2007) by removing features with low relevance towards the model's performance (Pelegrina and Siraj 2024), ensemble construction (Rozemberczki and Sarkar 2021), and the pruning of neural networks (Ghorbani and Zou 2020).

Further practical applications include its usage to quantify each feature's impact in predicting the risk degree in managing industrial machine maintenance (Nimmy et al. 2023), Pelegrina, Duarte, and Grabisch (2023b) apply it to evaluate the influence of each electrode on the quality of recovered fetal electrocardiograms, and Brusa et al. (2023) measure the features' importance towards machinery fault detection. Worth mentioning, each application requires an appropriate modeling in terms of player set and value function in order to obtain meaningful scores.

The uniqueness of the Shapley value comes at a price that poses an inherent drawback to practitioners: its computation scales exponentially with the number of players taking part in the cooperative game. Consequently, it becomes quickly infeasible for increasing feature numbers or even a few datapoints, especially when complex models are in use whose evaluation is highly resource consuming. As a viable remedy, it is common practice to approximate the Shapley value while providing reliably precise estimates is crucial to obtain meaningful importance scores. On this background, the recent interest in assigning scores to features, datapoints, or even model components, has fueled the research on approximation algorithms, leading to a diverse landscape of approaches (see (Chen et al. 2023) for an overview).

Contribution. We propose with *SVA*k_{ADD} (Shapley Value Approximation under k-additivity) a novel approximation method for the Shapley value based on the concept of k-additive games whose structure elicits a denser parameterizable value function. Fitting a k-additive surrogate game to randomly sampled coalition-value pairs comes with a twofold benefit. First, it reduces flexibility, promising faster convergence and second, the Shapley values of the k-additive surrogate game are obtained immediately from its representation. In summary, our contributions are:

- *SVA*k_{ADD} fits a k-additive surrogate game to sampled coalitions, mimicking the given game by a simpler structure with parameterizable degree of freedom while maintaining low representation error. The surrogate game's own Shapley values are obtained immediately and yield precise estimates for the given game if the representation exhibits a good fit.
- *SVA*k_{ADD} does not require any structural properties of the value function. Thus, it is domain-independent and can be applied to any cooperative game oblivious to what players and payoffs represent. Specifically in the field of explainability, it is model-agnostic and can approximate local as well as global explanations.
- We prove the theoretical soundness of *SVA*k_{ADD} by showing analytically that its underlying optimization problem yields the Shapley value.
- We empirically compare *SVA*k_{ADD} to competitive baselines at the hand of various explanation tasks, and shed light onto the best fitting degree of k-additivity.

2 Related Work

The problem of approximating the Shapley value, tackled by various communities, lead to a multitude of approaches to overcome its complexity. First to mention among the class of methods that can handle arbitrary games, without further assumptions on the value function, are those which construct mean estimates via random sampling. Fittingly, the Shapley value of a player can be interpreted as its expected marginal contribution to a specific probability distribution over coalitions. Castro, Gómez, and Tejada (2009) propose with *ApproShapley* the sampling of permutations from which marginal contributions are extracted. Further works employ stratification by coalition size (Maleki et al. 2013; Castro et al. 2017; van Campen et al. 2018; Okhrati and Lipani 2020; Zhang et al. 2023), or utilize reproducing kernel Hilbert spaces (Mitchell et al. 2022). Departing from marginal contributions, *Stratified SVARM* (Kolpaczki et al. 2024a) splits the Shapley value into multiple means of coalition values being further refined by *Adaptive SVARM* (Kolpaczki, Haselbeck, and Hüllermeier 2024). Guided by a different representation of the Shapley value, *KernelSHAP* (Lundberg and Lee 2017) solves an approximated weighted least squares problem, to which the Shapley value is its solution. Fumagalli et al. (2023) prove its variant *Unbiased KernelSHAP* (Covert and Lee 2021) to be equivalent to importance sampling of single coalitions. Joining this family, Pelegrina, Duarte, and Grabisch (2023a) propose k_{ADD}-SHAP, which formulates the surrogate model assuming a k-additive game[1]. It locally adopts the Choquet integral as the interpretable model, whose parameters have a straightforward connection with the Shapley value. Similar to us, Yan, Kroer, and Peysakhovich (2020) apply surrogate games of parameterizable structure that sum up unanimity games to calculate Shapley values, however, under the assumption that the given game possesses this structure from which we refrain.

On the contrary, tailoring the approximation to a specific application by leveraging structural properties promises faster converging estimates. In data valuation, including knowledge of how datapoints tend to contribute to a learning algorithm's performance resulted in multiple tailored methods (Ghorbani and Zou 2019; Jia et al. 2019b,a). In similar fashion Liben-Nowell et al. (2012) leverage supermodularity in cooperative games. Even further, value functions of certain parameterized shapes facilitate closed-form polynomial solutions of the Shapley value w.r.t. the number of players. Examples include the voting game (Bilbao et al. 2000) and the minimum cost spanning tree games (Granot, Kuipers, and Chopra 2002) used in operations research.

3 The Shapley Value and k-Additivity

We formally introduce cooperative games and the Shapley value in Section 3.1. Next, we present in Section 3.2 the concept of k-additivity, constituting the core of our approach.

[1]Note that k_{ADD}-SHAP is limited to local explanations. In contrast, our proposed method *SVA*k_{ADD} differs by its applicability to any formulation of a cooperative game. Moreover, in the context of explainable AI, it is capable of providing global explanations.

3.1 Cooperative Games and the Shapley Value

A cooperative game is formally described by n players, captured by the set $N = \{1, \ldots, n\}$, and an associated payoff function $\nu : \mathcal{P}(N) \to \mathbb{R}$, where $\mathcal{P}(N)$ represents the power set of N. This simple but expressive formalism may for example represent a shipment coordination where companies form a coalition in order to save costs when delivering their products. In this case, the companies can be modeled as players and $\nu(A)$ represents the benefit achieved by the group of companies $A \subseteq N$. Clearly, $\nu(N)$ is the total benefit when all companies (players) form the grand coalition N. Commonly, one normalizes the game by defining $\nu(\emptyset) = 0$, i.e., the worth of the empty set. However, in explainability, $\nu(\emptyset)$ may take nonzero values, e.g., with no features available one may obtain a classification accuracy of 50%. In this case, one can normalize ν by simply subtracting the worth of the empty set from all game payoffs, i.e., $\nu'(A) \leftarrow \nu(A) - \nu(\emptyset)$ for all $A \subseteq N$.

A central question arising from a cooperative game is how to fairly share the worth $\nu(N)$ of the grand coalition N among all participating players. The Shapley value (Shapley 1953) emerges as the prevalent solution concept since it uniquely satisfies axioms that intuitively capture fairness (Shapley 1953). Given the game (N, ν), the Shapley value of each player i is defined as

$$\phi_i = \sum_{A \subseteq N \setminus \{i\}} \frac{(n - |A| - 1)!\,|A|!}{n!} [\nu(A \cup \{i\}) - \nu(A)]\,, \quad (1)$$

where $|A|$ represents the cardinality of coalition A. It can be interpreted as a player's weighted average of marginal contributions to the payoff. Among the fulfilled axioms such as null player, symmetry, and additivity (see (Young 1985) for more details and other properties), in explainability the most useful is efficiency. It demands that the sum of all players' Shapley values is equal to the difference between $\nu(N)$ and $\nu(\emptyset)$. Formally, efficiency means $\sum_{i=1}^{n} \phi_i = \nu(N) - \nu(\emptyset)$. Or, in the game theory framework where $\nu(\emptyset) = 0$, one obtains $\sum_{i=1}^{n} \phi_i = \nu(N)$. In explainability, efficiency can be used to decompose a measure of interest among the set of features. As a result, one can interpret the importance of each feature to that measure.

Unfortunately, satisfying the desired axioms in the form of the Shapley value comes at a price. According to Equation (1), the calculation requires the evaluation of all 2^n coalitions within the exponentially growing power set of N. In fact, the exact computation of the Shapley value is known to be NP-hard (Deng and Papadimitriou 1994). Hence, its exact computation does not only become practically infeasible for growing player numbers but it is also of interest that the evaluation of only a few coalitions suffices to retrieve precise estimates. For instance, a model has to be costly retrained and re-evaluated on a test dataset for each coalition if one is interested in the features' impact on the generalization performance. Therefore, a common goal is to approximate all Shapley values $\phi = (\phi_1, \ldots, \phi_n)$ of a given game (N, ν) by observing only a subset of evaluated coalitions $\mathcal{M} \subseteq \mathcal{P}(N)$. We denote the size of $\mathcal{M}$ by $T \in \mathbb{N}$ and refer to it as the available budget representing the number of samples an approximation algorithm is allowed to draw. The mean squared error (MSE) serves as a popular measure to quantify the quality of the obtained estimates $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_n)$ and is to be minimized: $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(\hat{\phi}_i - \phi_i)^2]$, where the expectation is taken w.r.t. the (potential) randomness of the approximation strategy.

3.2 Interaction Indices and k-Additivity

The underlying idea of measuring the impact (or share) of a single player i by means of its marginal contributions finds its natural extension to sets of players S in the Shapley interaction index (Murofushi and Soneda 1993; Grabisch 1997) by generalizing from marginal contributions to discrete derivatives. For any $S \subseteq N$ its Shapley interaction $I(S)$ is given by

$$I(S) = \sum_{A \subseteq N \setminus S} w_{A,S} \left(\sum_{A' \subseteq S} (-1)^{|S| - |A'|}\, \nu(A \cup A') \right) \quad (2)$$

with weights $w_{A,S} = \frac{(n - |A| - |S|)!|A|!}{(n - |S| + 1)!}$. For convenience, we will write $I_i := I(\{i\})$ and $I_{i,j} := I(\{i, j\})$. Instead of individual importance, $I(S)$ indicates the synergy between players in S. Although this interpretation is not straightforward for coalitions of three or more entities, it has a clear meaning for pairs. For two players i and j, the Shapley interaction index $I_{i,j}$ quantifies how the presence of i impacts the marginal contributions of j and vice versa. Especially in explainable AI, where players represent features, it can be interpreted as follows: (i) if $I_{i,j} < 0$, there is a negative interaction (redundant effect) between features i, j; (ii) if $I_{i,j} > 0$, there is a positive interaction (complementary effect) between features i, j; (iii) if $I_{i,j} = 0$, there is no interaction between i, j (independence) on average.

Note that the Shapley interaction index reduces to the Shapley value for a singleton, i.e., $I_i = \phi_i$. Moreover, there is a linear relation between the interactions and the game payoffs (Grabisch 1997). Indeed, from the interactions one may easily retrieve the game payoffs by the following expression:

$$\nu(A) = \sum_{B \subseteq N} \gamma_{|A \cap B|}^{|B|} I(B)\,, \quad (3)$$

where $\gamma_{|A \cap B|}^{|B|}$ is defined as follows with the Bernoulli numbers starting at $\eta_0 = 1$:

$$\gamma_r^s = \sum_{l=0}^{r} \binom{r}{l} \eta_{s-l} \quad \text{and} \quad \eta_r = -\sum_{l=0}^{r-1} \frac{\eta_l}{r - l + 1} \binom{r}{l}\,.$$

This linear transformation recovers any coalition value $\nu(A)$ by using the Shapley interactions of all 2^n coalitions, thus including the Shapley values. Therefore, 2^n parameters are to be defined if the whole game is to be expressed by Shapley interactions. However, in some situations one may assume that interactions only exist for coalitions up to k many players. This assumption leads to the concept known as k-additive games. A k-additive game is such that $I(S) = 0$ for all S with $|S| > k$. Obviously, this restricts the flexibility of

the game but depending on k, this may significantly decrease the number of parameters to be defined such that for low k it increases only polynomially with the number of players. For instance, in 2-additive and 3-additive games, there are only $n(n+1)/2$, and $n(n^2+5)/6$ respectively, many interactions indices as the remaining parameters are equal to zero. One may argue that within Shapley-based feature explanations, the neglect of higher order interactions, by setting them to zero per default, comes naturally. For instance, (Bordt and von Luxburg 2023) show that these interactions barely exist in the context of post-hoc local explanations.

4 k-Additive Approximation Approach

In this section, we present our method *SVA*k_{ADD} to approximate Shapley values. It builds upon the idea of adjusting a k-additive surrogate game (N, ν_k) to randomly sampled and evaluated coalitions. Having fit the surrogate game to represent the observed coalition values with minimal error, its own Shapley values ϕ^k can be interpreted as estimates $\hat{\phi}$ for ϕ of (N, ν) since the fitting promises ν_k to be close to ν. See Figure 1 for an illustration of the approach. The framework of fitting a surrogate game encompasses other methods such as *KernelSHAP* (Lundberg and Lee 2017) and k_{ADD}-SHAP (Pelegrina, Duarte, and Grabisch 2023a). See Appendix C for a conceptual comparison with our proposal.

4.1 The k-Additive Optimization Problem

We leverage the representation of ν_k by means of interactions as given in Equation (3). In particular, since ν_k is supposed to be k-additive, we specify ν_k as a linear transformation of interactions $I^k(B)$ for all $B \subseteq N$ of size $|B| \leq k$, allowing us to truncate interactions of higher order than k:

$$\nu_k(A) = \sum_{B \subseteq N, |B| \leq k} \gamma_{|A \cap B|}^{|B|} I^k(B)\,. \tag{4}$$

Within this representation, the Shapley values ϕ^k of the resulting game (N, ν_k) are obtained immediately by the interactions $I_i^k = \phi_i^k$, which will serve as estimates for the Shapley values ϕ of the game (N, ν), i.e. $I_i^k \approx \phi_i$. The k-additive representation of ν_k comes with the advantage that the number of parameters $I^k(B)$ needed to define the surrogate game is reduced (as several parameters are set to zero). The drawback of this strategy is the reduction in flexibility left to model the observed game (N, ν) according to the obtained evaluations. However, we can still model interactions for coalitions up to k players. Empirically, works in the literature (Grabisch et al. 2002, 2006; Pelegrina et al. 2020; Pelegrina, Duarte, and Grabisch 2023a) have been using 2-additive or even 3-additive games and obtained satisfactory results for modeling interactions. Our goal is to fit ν_k as closely as possible to ν and therefore minimize the deviation of ν_k from ν captured by

$$\sum_{A \in \mathcal{P}(N) \setminus \{\emptyset, N\}} w_A \left(\nu(A) - \nu_k(A)\right)^2\,, \tag{5}$$

where w_A is an importance weight associated to each coalition A. We are eager to meet the desirable efficiency axiom

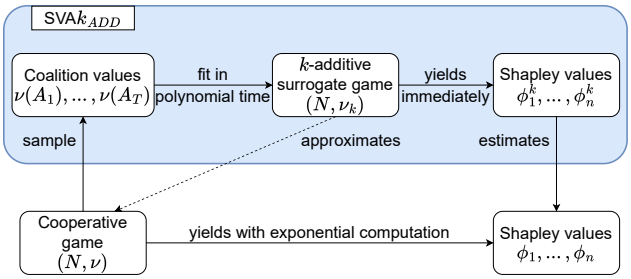


Figure 1: The sampled coalition values $\nu(A_1), \ldots, \nu(A_T)$ from the game (N, ν) are used to fit a k-additive surrogate game (N, ν_k) in polynomial time w.r.t. n. The Shapley values $\phi_1^k, \ldots, \phi_n^k$ of (N, ν_k) are obtained immediately from its k-additive representation. As ν_k approximates ν, these serve as estimates of the true Shapley values $\phi_1, \ldots, \phi_n$ of (N, ν).

such that the difference between and $\nu(N)$ and $\nu(\emptyset)$ is decomposed among the players within our approximated values ϕ^k. This is ensured by imposing the constraint $\nu(N) - \nu(\emptyset) = \nu_k(N) - \nu_k(\emptyset)$. Hence, we arrive at the following optimization problem.

Definition 4.1. Given a cooperative game (N, ν), a degree of k-additivity $k \in \mathbb{N}$ with $k \leq n$, and weights $w_A \in \mathbb{R}$ associated with each coalition $A \subseteq N$, the k-additive optimization problem is given by the following constrained weighted least square optimization problem:

$$\begin{aligned} \min_{I^k} \quad & \sum_{A \in \mathcal{P}(N) \setminus \{\emptyset, N\}} w_A \left(\nu(A) - \sum_{B \subseteq N, |B| \leq k} \gamma_{|A \cap B|}^{|B|} I^k(B)\right)^2 \\ \text{s.t.} \quad & \nu(N) - \nu(\emptyset) = \sum_{B \subseteq N, |B| \leq k} \left(\gamma_{|B|}^{|B|} - \gamma_0^{|B|}\right) I^k(B) \end{aligned}$$

Solving the k-additive optimization is at the core of our approach. In the remainder we describe how to overcome two key challenges. First, we address in Section 4.2 how to choose the weights w_A such that ϕ^k comes close to ϕ. Second, as the objective function sums up over exponential many coalitions, we present in Section 4.3 our algorithm *SVA*k_{ADD} that constructs an approximative objective function by sampling coalitions and adding their error terms.

4.2 Theoretical Soundness Through Choice of Weights

Seeking precise estimates $\phi^k \approx \phi$, one may even raise the question if it is feasible to retrieve the exact Shapley values ϕ from the solution I^k and how the weights w_A have to be set to achieve this. We analytically derive the correct weights and positively answer this question.

Theorem 4.2. *The solution to the k-additive optimization problem of any cooperative game (N, ν) for the cases of $k = 1$, $k = 2$, and $k = 3$ with weights $w_A^* = \binom{n-2}{|A|-1}^{-1}$ yields the Shapley value, i.e.*

$$I_i^k = \phi_i\,.$$

See Appendix A for the proof of Theorem 4.2. Note that the weights coincide with those derived by (Charnes et al.

1988) used in (Lundberg and Lee 2017) for a different optimization problem. The result implies that after observing all coalitions of the cooperative game (N, ν) our approach yields the exact Shapley values with no approximation error. We interpret this as evidence for the soundness and theoretical foundation of our method. Moreover, since the result holds irregardless of the shape of ν, the game can even highly deviate from being k-additive and our estimates will still converge to ϕ. Hence, k-additivity is not an assumption that our method requires but rather a tool to be leveraged. We conjecture that Theorem 4.2 holds also true for arbitrary degrees of k-additivity and leave the proof for future work due to the analytical challenges. Worth mentioning is that the hardness of incorporating Shapley interactions of higher degree into weighted least squares optimizations has been acknowledged by (Fumagalli et al. 2024).

4.3 Approximating the k-Additive Optimization Problem via Sampling

Computing the solution to the k-additive optimization problem (see Definition 4.1) is practically infeasible since the objective compromises exponential many error terms w.r.t. n. As a remedy we follow the same strategy as adopted in (Lundberg and Lee 2017; Pelegrina, Duarte, and Grabisch 2023a) and approximate the objective function by sampling coalitions without replacement. Let $\mathcal{M} = \{A_1, \ldots, A_T\}$ be the set of sampled coalitions with $A_i \neq A_j$ for all $i \neq j$ and the sequence $\nu_{\mathcal{M}} = (\nu(A_1), \ldots, \nu(A_T))$ representing its evaluated coalition values. Thus, we solve the following optimization problem after sampling:

$$\begin{aligned} \min_{I^k} \quad & \sum_{A \in \mathcal{M} \setminus \{\emptyset, N\}} w_A \left(\nu(A) - \sum_{B \subseteq N, |B| \leq k} \gamma_{|A \cap B|}^{|B|} I^k(B) \right)^2 \\ \text{s.t.} \quad & \nu(N) - \nu(\emptyset) = \sum_{B \subseteq N, |B| \leq k} \left(\gamma_{|B|}^{|B|} - \gamma_0^{|B|} \right) I^k(B) \end{aligned} \tag{6}$$

To ensure the efficiency constraint, we force the evaluation of $\nu(\emptyset)$ and $\nu(N)$. Each coalition $A \in \mathcal{P}(N) \setminus \{\emptyset, N\}$ is drawn according to an initial probability distribution p defined by $p_A = \frac{w_A^*}{\sum_{B \in \mathcal{P}(N) \setminus \{\emptyset, N\}} w_B^*}$ (see Appendix B.2 for a practical realization of the sampling). After drawing a coalition A, we set p_A to zero and normalize the remaining probabilities. This procedure is repeated until $|\mathcal{M}| = T$. Algorithm 1 presents the pseudo-code of *SVA*k_{ADD}. The algorithm requires the game (N, ν), the additivity degree k, and the budget T. It starts by evaluating $\nu(\emptyset)$ and $\nu(N)$. Thereafter, based on the (normalized) distribution p, it samples $T - 2$ coalitions from $\mathcal{P}(N) \setminus \{\emptyset, N\}$, evaluates each, and extends $\mathcal{M}$ as well as $\nu_{\mathcal{M}}$. Finally, it solves the optimization problem in Equation (6) with weights w_A^* given by Theorem 4.2 (see Appendix B.1 for an analytical solution). The extracted Shapley values ϕ^k of ν_k are returned as estimates $\hat{\phi}$ for the Shapley values ϕ of (N, ν).

We would like to emphasize that Theorem 4.2 does not make a statement about I_i^k during sampling when not all coalitions are observed. To the best of our knowledge, there exists no approximation guarantee for methods that estimate the Shapley value by means of a weighted least squares optimization problem. The difficulty of obtaining a theoretical result is further elaborated by (Covert and Lee 2021).

Algorithm 1: *SVA*k_{ADD}

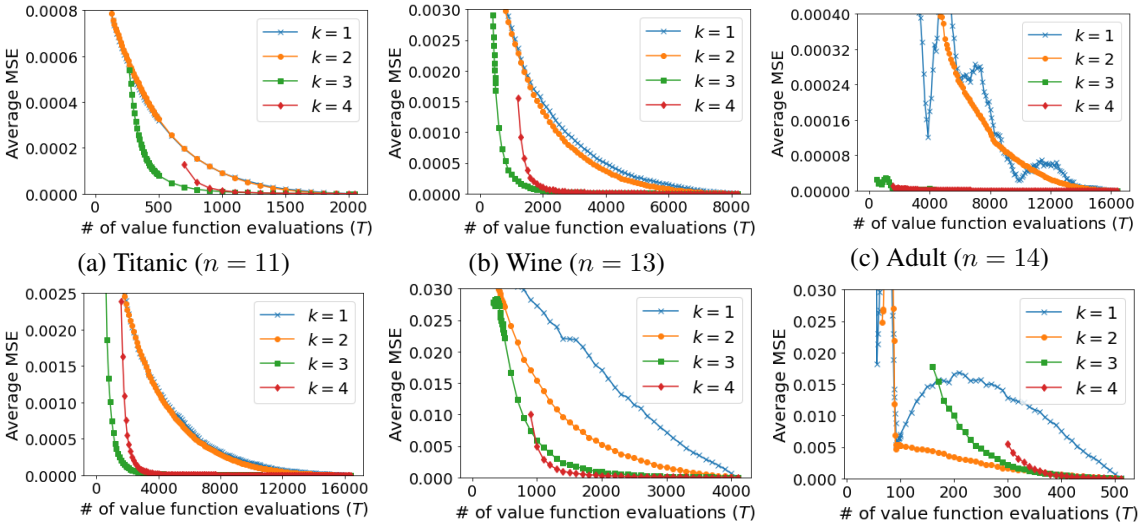
Input: (N, ν), k, T
$\mathcal{M} \leftarrow \{\emptyset, N\}$
$\nu_{\mathcal{M}} \leftarrow (\nu(\emptyset), \nu(N))$
while $|\mathcal{M}| < T$ **do**
  Sample a coalition $A \in \mathcal{P}(N) \setminus \{\emptyset, N\}$ from normalized distribution p
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{A\}$
  $\nu_{\mathcal{M}} \leftarrow (\nu_{\mathcal{M}}, \nu(A))$
  $p_A \leftarrow 0$
end while
$(I^k(B))_{B \subseteq N : |B| \leq k} \leftarrow \text{SOLVE}(\mathcal{M}, \nu_{\mathcal{M}}, k)$
Output: $I_1^k, \ldots, I_n^k$

In addition, we apply a modified version of border sampling utilized by (Fumagalli et al. 2023; Kolpaczki et al. 2024b; Lundberg and Lee 2017) that demonstrates empirical improvements. Since the coalition sizes at the ends of the spectrum close to 0 or n comprise relatively few coalitions, we first deterministically collect all coalitions of size 1,2,$n-2$, and $n-1$ and then continue with the random sampling of coalitions with cardinality between 3 and $n-3$.

5 Empirical Evaluation

In order to assess the approximation performance of *SVA*k_{ADD}, we conduct experiments with cooperative games stemming from various explanation types. Although our method is not limited to a certain domain, we find approximating feature scores best to exemplarily illustrate its effectiveness on several datasets as a sanity check. Our evaluation is mainly two-fold. Not only are we interested in the comparison of *SVA*k_{ADD} against current state-of-the-art model-agnostic methods in Section 5.2, but we also seek to investigate how the choice of the assumed degree of additivity k affects the approximation quality (see Section 5.3). In the sequel of Section 5.1, we describe the utilized datasets and resulting cooperative games.

For each considered combination of dataset, approximation algorithm, and budget T, the obtained estimates $\hat{\phi}$ are compared with ϕ which we calculate exhaustively in advance. We measure approximation quality of the estimates by the MSE. The error is measured depending on T as we intentionally refrain from a runtime comparison for multiple reasons: (i) the observed runtimes may differ depending on the actual implementation, (ii) evaluating the worth of a coalition poses the bottleneck in explanation tasks, rendering the difference in performed arithmetic operations negligible for more complex models and datasets, (iii) instead of runtime, monetary units might be paid for each access to a remotely provided model offered by a third-party.



(a) Titanic ($n = 11$) (b) Wine ($n = 13$) (c) Adult ($n = 14$)

(d) ImageNet ($n = 14$) (e) Big Five ($n = 12$) (f) Breast Cancer ($n = 9$)

Figure 2: MSE of $SVAk_{\text{ADD}}$, averaged over 50 runs except for (c) and (d) with 10 runs, in dependence of available budget T for different additivity degrees k. Datasets stem from various explanation types: global (a)-(b), local (c)-(d), and unsupervised (e)-(f) with differing player numbers n.

5.1 Datasets

We distinguish between three feature explanation tasks: global importance, local attribution, and unsupervised importance being described further in **??**.

Within global importance (Covert, Lundberg, and Lee 2020) the features' contributions to a model's generalization performance are quantified. This is done by means of accuracy for classification and the MSE for regression on a test set. For each evaluated coalition a random forest is trained on a training set. We employ the *Titanic* (classification, 11 features) and *Wine* dataset (classification, 13 features).

On the contrary, local feature attribution (Lundberg and Lee 2017) measures each feature's impact on the prediction of a fixed model for a given datapoint. While the predicted value can directly be used as the worth of a feature coalition for regression, the predicted class probability is required instead of a label for classification. Rendering a feature outside of an evaluated coalition absent is performed by means of imputation that blurs the features contained information. The experiments are conducted on the *Adult* (classification, 14 features) and *ImageNet* (classification, 14 features) dataset.

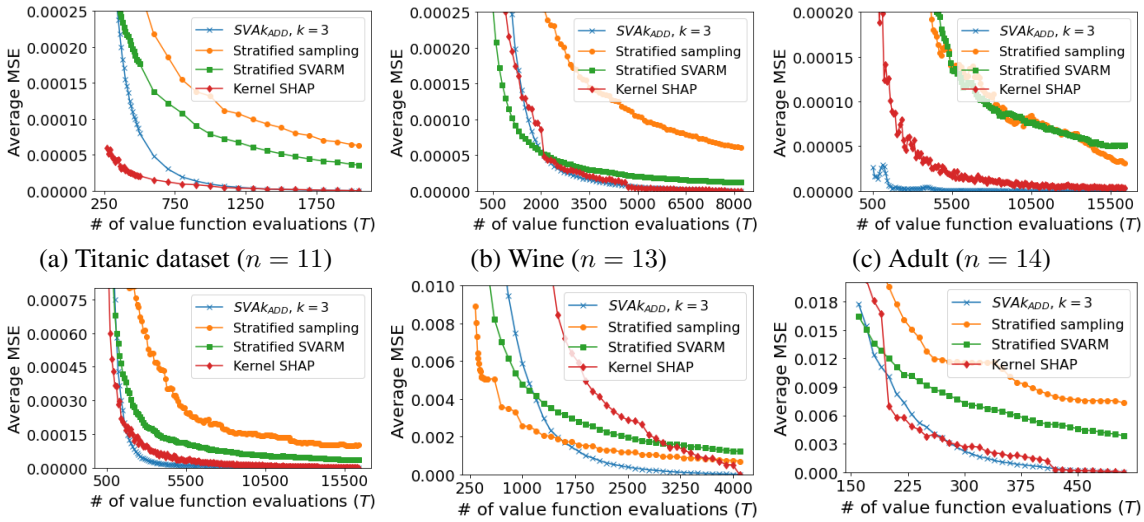
In the absence of labels, unsupervised feature importance (Balestra et al. 2022) seeks to find scores without a model's predictions. This is achieved by employing the total correlation of a feature subset as its worth, since the datapoints can be seen as realizations of the joint feature value distribution. For this task, we consider the *Big Five* (12 features) and *Breast Cancer* (9 features) datasets.

5.2 Impact of the Additivity Degree k

In order to provide an understanding of the underlying trade-off between fast convergence (low k) and expressiveness (high k) of the surrogate game and how the crucial choice of k affects the approximation quality, we evaluate $SVAk_{\text{ADD}}$ for different $k \in \{1, 2, 3, 4\}$ in Figure 2. See **??** for more results with the *Diabetes*, *ImageNet*, and *Breast cancer* dataset.

The curves for higher k begin at points of higher budget because the greater k, the more coalition values are required to identify a unique k-additive value function that fits the observations. We explain the behavior for low k, specifically $k = 1$ and $k = 2$, by the model's inability to achieve a good fit due to missing flexibility. As a result, the convergence to the exact Shapley values is slow. These findings imply that interactions up to order 2 are not sufficient to model how features jointly impact performance (global task) or prediction outcome (local task). On the other hand, both the 3-additive and 4-additive model converge significantly faster for most datasets and outperform the parameterization with $k = 1$ or $k = 2$ after a few samples. The choice of $k = 3$ appears preferable as it results in quicker decreasing error curves.

Choice of k. The problem of choosing the optimal k is intriguing and could potentially lead to further fruitful research. It is comparable to related problems in other domains, for example finding the best k in k-means clustering, and methods used there may also apply to our case. A well-known example is the elbow heuristic: One fits multiple k-additive surrogate games, with increasing value for k, and monitors the resulting approximation error. To our advantage, these multiple games can be fitted with one single stream of sampled coalitions, and hence we do not reduce the effectively available budget T. The fit must improve monotonically for increasing k (just like the objective function in k-means can only decrease with increasing k), because a larger k only adds additional free variables in the optimization problem. However, one typically observes



(a) Titanic dataset ($n = 11$) (b) Wine ($n = 13$) (c) Adult ($n = 14$)

(d) ImageNet ($n = 14$) (e) Big Five ($n = 12$) (f) Breast Cancer ($n = 9$)

Figure 3: MSE of $SVAk_{\text{ADD}}$ and competing methods, averaged over 50 runs except for (c) and (d) with 10 runs, in dependence of available budget T. Datasets stem from various explanation types: global (a)-(b), local (c)-(d), and unsupervised (e)-(f) with differing player numbers n.

bigger gains in the beginning (where k is small), which at some point start to become much smaller. Beyond this point (the "elbow"), the additional gains by increasing k are only marginal, so the elbow determines a good value for k.

5.3 Comparison with Existing Methods

In our second experiment, we compare $SVAk_{\text{ADD}}$ with other existing approximation methods. For instance, we consider *Stratified sampling* (Maleki et al. 2013), *Stratified SVARM* (Kolpaczki et al. 2024a) and *KernelSHAP* (Lundberg and Lee 2017). For the purpose of comparison, we adopt the 3-additive model to represent $SVAk_{\text{ADD}}$ since it displays the most satisfying compromise between approximation quality and minimum required evaluations as argued in Section 5.2. Figure 3 presents the obtained results for all methods. See **??** for more results with the *Diabetes*, *IMDB* and *FIFA 21* dataset, including *Permutation sampling* (Castro, Gómez, and Tejada 2009) and the 2-additive model.

First to mention is that $SVAk_{\text{ADD}}$ competes consistently with *Stratified SVARM* and *KernelSHAP* for the best approximation performance across all datasets. Although for a very low number of function evaluations $SVAk_{\text{ADD}}$ achieves an error greater than some other approaches, at some point during the approximation process it converges faster to the exact Shapley values and leaves it competitors with a considerable margin behind, especially for local feature attribution on the *Adult* and *ImageNet* dataset. In comparison to *KernelSHAP* our method $SVAk_{\text{ADD}}$ converges faster to the exact Shaley values for the *Adult*, *Big Five*, *ImageNet*, and *Wine* dataset, whereas for the *Titanic* and *Breast Cancer* dataset, *KernelSHAP* achieves a better performance to which $SVAk_{\text{ADD}}$ catches up with sufficient budget.

6 Conclusion

We proposed with $SVAk_{\text{ADD}}$ a new algorithm to approximate Shapley values that fits a structured surrogate game instead of providing mean estimates via Monte Carlo sampling. Despite restricting the surrogate game to be k-additive, our developed method is model-agnostic. It is also applicable to any cooperative game without posing further assumptions since its underlying optimization problem provably yields the Shapley value. We investigated empirically the trade-off that the choice of the parameter k poses. Further, $SVAk_{\text{ADD}}$ exhibits competitive results with other existing approaches depending on the considered explanation type, dataset, and available budget, allowing us to conclude the non-existence of a dominating approximation method.

Limitations and future work. While the surrogate game's flexibility increases with higher k-additivity, it also requires more observations to obtain a unique solution of the optimization problem, eventually posing a practical limit on k. The choice of k is seemingly non-trivial and employing the elbow heuristic could provide a potential solution. As the fit must improve monotonically for increasing k, one could fit multiple k-additive surrogate games and select the k at the elbow where the gains in fit start to become significantly smaller. We expect future investigations of differently structured surrogate games to yield likewise fruitful results and contribute to the advancement of this class of approximation algorithms. Besides the estimated Shapley values, our proposal could also provide interaction effects when $k \geq 2$. Although we did not address these parameters, future works can extract the estimated interactions to investigate redundant or complementary features. This could be of interest in practical applications where interaction between features are relevant as for example in disease detection.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP, grant number 2025/00700-0) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, TRR 318/1 2021 – 438445824).

References

Balestra, C.; Huber, F.; Mayr, A.; and Müller, E. 2022. Unsupervised Features Ranking via Coalitional Game Theory for Categorical Data. In *Proc. of Big Data Analytics and Knowledge Discovery (DaWaK)*, 97–111.

Bilbao, J.; Fernández, J.; Jiménez-Losada, A.; and López, J. 2000. Generating Functions for Computing Power Indices Efficiently. *Top*, 8: 191–213.

Bordt, S.; and von Luxburg, U. 2023. From Shapley Values to Generalized Additive Models and back. In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 709–745.

Brusa, E.; Cibrario, L.; Delprete, C.; and Di Maggio, L. G. 2023. Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring. *Applied Sciences (Switzerland)*, 13(4).

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

Castro, J.; Gómez, D.; and Tejada, J. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5): 1726–1730.

Charnes, A.; Golany, B.; Keane, M.; and Rousseau, J. 1988. *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, 123–133. Springer Netherlands.

Chen, H.; Covert, I. C.; Lundberg, S. M.; and Lee, S. 2023. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5(6): 590–601.

Cohen, S. B.; Dror, G.; and Ruppin, E. 2007. Feature Selection via Coalitional Game Theory. *Neural Comput.*, 19(7): 1939–1961.

Cohen, S. B.; Ruppin, E.; and Dror, G. 2005. Feature Selection Based on the Shapley Value. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 665–670.

Covert, I.; and Lee, S.-I. 2021. Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3457–3465.

Covert, I.; Lundberg, S. M.; and Lee, S. 2020. Understanding Global Feature Contributions With Additive Importance Measures. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Deng, X.; and Papadimitriou, C. H. 1994. On the Complexity of Cooperative Solution Concepts. *Math. Oper. Res.*, 19(2): 257–266.

Fiestras-Janeiro, M. G.; García-Jurado, I.; Meca, A.; and Mosquera, M. A. 2011. Cooperative game theory and inventory management. *European Journal of Operational Research*, 210: 459–466.

Fumagalli, F.; Muschalik, M.; Kolpaczki, P.; Hüllermeier, E.; and Hammer, B. 2023. SHAP-IQ: Unified Approximation of any-order Shapley Interactions. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Fumagalli, F.; Muschalik, M.; Kolpaczki, P.; Hüllermeier, E.; and Hammer, B. 2024. KernelSHAP-IQ: Weighted Least Square Optimization for Shapley Interactions. In *Proc. of the 41st International Conference on Machine Learning (ICML)*.

Ghorbani, A.; and Zou, J. Y. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proc. of the 36th International Conference on Machine Learning (ICML)*, volume 97, 2242–2251.

Ghorbani, A.; and Zou, J. Y. 2020. Neuron Shapley: Discovering the Responsible Neurons. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Grabisch, M. 1997. Alternative representations of discrete fuzzy measures for decision making. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 5: 587–607.

Grabisch, M.; Duchêne, J.; Lino, F.; and Perny, P. 2002. Subjective evaluation of discomfort in sitting positions. *Fuzzy Optimization and Decision Making*, 1: 287–312.

Grabisch, M.; Prade, H.; Raufaste, E.; and Terrier, P. 2006. Application of the Choquet integral to subjective mental workload evaluation. *IFAC Proceedings Volumes*, 39: 135–140.

Granot, D.; Kuipers, J.; and Chopra, S. 2002. Cost Allocation for a Tree Network with Heterogeneous Customers. *Mathematics of Operations Research*, 27(4): 647–661.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Gürel, N. M.; Li, B.; Zhang, C.; Spanos, C. J.; and Song, D. 2019a. Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms. *Proc. VLDB Endow.*, 12(11): 1610–1623.

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019b. Towards Efficient Data Valuation Based on the Shapley Value. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1167–1176.

Kolpaczki, P.; Bengs, V.; Muschalik, M.; and Hüllermeier, E. 2024a. Approximating the Shapley Value without Marginal Contributions. In *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*, 13246–13255.

Kolpaczki, P.; Haselbeck, G.; and Hüllermeier, E. 2024. How Much Can Stratification Improve the Approximation of Shapley Values? In *Proc. of World Conference on Explainable Artificial Intelligence (xAI)*, 489–512.

Kolpaczki, P.; Muschalik, M.; Fumagalli, F.; Hammer, B.; and Hüllermeier, E. 2024b. SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification. In *The 27nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3520–3528.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9): 1–46.

Liben-Nowell, D.; Sharp, A.; Wexler, T.; and Woods, K. M. 2012. Computing Shapley Value in Supermodular Coalitional Games. In *Computing and Combinatorics - 18th Annual International Conference COCOON*, 568–579.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 4768–4777.

Maleki, S.; Tran-Thanh, L.; Hines, G.; Rahwan, T.; and Rogers, A. 2013. Bounding the Estimation Error of Sampling-based Shapley Value Approximation With/Without Stratifying. *CoRR*, abs/1306.4265.

Marcílio, W. E.; and Eler, D. M. 2020. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347.

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

Molnar, C. 2021. *Interpretable machine learning*.

Murofushi, T.; and Soneda, S. 1993. Techniques for reading fuzzy measures (III): interaction index. In *9th fuzzy system symposium*, 693–696.

Nimmy, S. F.; Hussain, O. K.; Chakrabortty, R. K.; Hussain, F. K.; and Saberi, M. 2023. Interpreting the antecedents of a predicted output by capturing the interdependencies among the system features and their evolution over time. *Engineering Applications of Artificial Intelligence*, 117(November 2022): 105596.

Okhrati, R.; and Lipani, A. 2020. A Multilinear Sampling Algorithm to Estimate Shapley Values. In *25th International Conference on Pattern Recognition ICPR*, 7992–7999.

Pelegrina, G. D.; Duarte, L. T.; and Grabisch, M. 2023a. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artificial Intelligence*, 325: 104014.

Pelegrina, G. D.; Duarte, L. T.; and Grabisch, M. 2023b. Interpreting the contribution of sensors in blind source extraction by means of Shapley values. *IEEE Signal Processing Letters*, 30(1): 878–882.

Pelegrina, G. D.; Duarte, L. T.; Grabisch, M.; and Romano, J. M. T. 2020. The multilinear model in multicriteria decision making: The case of 2-additive capacities and contributions to parameter identification. *European Journal of Operational Research*, 282.

Pelegrina, G. D.; and Siraj, S. 2024. Shapley value-based approaches to explain the quality of predictions by classifiers. *IEEE Transactions on Artificial Intelligence*, 1–15.

Pfannschmidt, K.; Hüllermeier, E.; Held, S.; and Neiger, R. 2016. Evaluating Tests in Medical Diagnosis: Combining Machine Learning with Game-Theoretical Concepts. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume 610 of *Communications in Computer and Information Science*, 450–461.

Rozemberczki, B.; and Sarkar, R. 2021. The Shapley Value of Classifiers in Ensemble Games. In *The 30th ACM International Conference on Information and Knowledge Management CIKM*, 1558–1567.

Rozemberczki, B.; Watson, L.; Bayer, P.; Yang, H.-T.; Kiss, O.; Nilsson, S.; and Sarkar, R. 2022. The Shapley Value in Machine Learning. In *Proc. of the 31th International Joint Conference on Artificial Intelligence (IJCAI)*, 5572–5579.

Shapley, L. S. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, 307–318. Princeton University Press.

van Campen, T.; Hamers, H.; Husslage, B.; and Lindelauf, R. 2018. A new approximation method for the Shapley value applied to the WTC 9/11 terrorist attack. *Social Network Analysis and Mining*, 8(3): 1–12.

Yan, T.; Kroer, C.; and Peysakhovich, A. 2020. Evaluating and Rewarding Teamwork Using Cooperative Game Abstractions. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*.

Young, H. P. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14: 65–72.

Zhang, J.; Sun, Q.; Liu, J.; Xiong, L.; Pei, J.; and Ren, K. 2023. Efficient Sampling Approaches to Shapley Value Approximation. *Proc. ACM Manag. Data*, 1(1): 48:1–48:24.