Approximating the Shapley Value and Shapley Interactions

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
Doctor rerum naturalium (Dr. rer. nat.)

eingereicht von

Patrick Irenäus Kolpaczki

am 8. August 2025

Patrick Irenäus Kolpaczki

Approximating the Shapley Value and Shapley Interactions

1. Gutachter: Prof. Dr. Eyke Hüllermeier
Ludwig-Maximilians-Universität München

2. Gutachter: Prof. Dr. Michel Grabisch
Université Paris 1 Panthéon-Sorbonne

3. Gutachter: Prof. Dr. Thomas Nagler
Ludwig-Maximilians-Universität München

Tag der Einreichung: 8. August 2025
Tag der Disputation: 10. Oktober 2025

Acknowledgement

To begin with, I would like to thank Eyke Hüllermeier for giving me the opportunity to become a doctoral student and pursue my passion. I have enjoyed a certain freedom while conducting my research, and I want to express my gratitude for the trust that I have received under his supervision, enabling me to work autonomously to such a degree that I would like to call a privilege. I would like to express my thanks to all members of my examination committee and their engagement with my work. I thank Reinhold Häb-Umbach for taking care of formalities.

I am thankful for the financial support I received from Paderborn University and the state of North Rhine-Westphalia. Being able to connect with other doctoral students within the graduate school DataNinja and even just the exchange about our shared struggles made me feel a part of a group and encouraged me to overcome those. I would like to thank Moritz Lange and Raphael Engelhardt for their joyful companionship during that time and at conferences.

I owe my coauthors my deepest thanks for their contributions to my work. Especially the harmonious collaboration with Maximilian Muschalik and Fabian Fumagalli, and attending conferences together has been a joy. I am grateful to have worked with the talented students Patrick Becker, Georg Haselbeck, and Tim Nielen whose efforts went beyond my expectations. My thanks also go to all my colleagues at the Ludwig Maximilian University for supporting me whenever I asked for help. Our talks during lunchtime have been amusing, to say the least.

Without Viktor Bengs I would not be where I am today. You offered me a student job, supervised my master's thesis, shaped my writing style, and taught me how to write papers. You stood with me at the whiteboard whenever I had some tricky math to solve or was just hunting constants. You were there when I needed to talk to someone. I was happy whenever I saw you in the office. Thank you!

Finally, I want to thank my family and friends on whom I could always rely. My friend Jakub for always listening to me and sharing his honest but also soothing life experience. Our conversations have been a pleasure. My friend Omar for all the running training, bike tours, and the contagious attitude of facing life, and whatever it throws at you, with a certain relentlessness inspired by endurance sports. My cousin Maciek for all the joyful time we have spent together, helping me to clear my mind. From the bottom of my heart I would like to thank my parents for their tireless efforts in raising me, supporting me, and preparing me for this journey.

Danksagung

An erster Stelle möchte ich Eyke Hüllermeier für die Möglichkeit danken, Doktorand zu werden und meiner Leidenschaft nachzugehen. Ich habe während meiner Forschung eine gewisse Freiheit genossen und will meine Dankbarkeit für das Vertrauen ausdrücken, das ich unter seiner Betreuung erhalten habe und mir ermöglicht hat, mit einem Grad an Selbstständigkeit zu arbeiten, den ich gerne als Privileg bezeichnen würde. Ich möchte allen Mitgliedern meiner Prüfungskommission und ihrer Bereitschaft, sich mit meiner Arbeit auseinanderzusetzen, danken. Ich bedanke mich bei Reinhold Häb-Umbach für das Übernehmen von Formalitäten.

Ich bin dankbar für die finanzielle Unterstützung, die ich von der Universität Paderborn und dem Land Nordrhein-Westfalen erhalten habe. Sich im Graduiertenkolleg Dataninja mit anderen Doktoranden vernetzen und über gemeinsame Schwierigkeiten austauschen zu können, gab mir das Gefühl, Teil einer Gruppe zu sein, und motivierte mich ebenjene zu überwinden. Ich möchte mich bei Moritz Lange und Raphael Engelhardt für ihre Gesellschaft während dieser Zeit und auf Konferenzen bedanken.

Ich schulde meinen Koautoren meinen tiefsten Dank für ihre Unterstützung. Insbesondere die harmonische Zusammenarbeit mit Maximilian Muschalik und Fabian Fumagalli und das gemeinsame Besuchen von Konferenzen bereiteten mir Freude. Ich schätze mich glücklich mit den talentierten Studenten Patrick Becker, Georg Haselbeck und Tim Nielen gearbeitet zu haben, deren Einsatz meine Erwartungen überstieg. Mein Dank richtet sich auch an meine Kollegen an der Ludwig-Maximilians-Universität, die mich unterstützt haben, wann immer ich nach Hilfe gefragt habe. Unsere Gespräche während der Mittagspause waren mehr als amüsant.

Ohne Viktor Bengs wäre ich heute nicht da, wo ich jetzt bin. Du hast mir eine Stelle als wissenschaftliche Hilfskraft angeboten, meine Masterarbeit betreut, meinen Schreibstil geprägt und mir beigebracht, wie man wissenschaftliche Arbeiten schreibt. Du standest mit mir an der Tafel, wenn ich knifflige Mathematikprobleme zu lösen hatte oder einfach nur auf Konstantenjagd war. Du warst für mich da, wenn ich jemanden zum Reden brauchte. Ich war jedes Mal glücklich, dich im Büro zu sehen. Danke!

Zu guter Letzt will ich mich bei meiner Familie und meinen Freunden bedanken, auf die ich mich stets verlassen konnte. Mein Freund Jakub, der mir immer sein Gehör geschenkt und seine ehrliche aber beruhigende Lebensweisheit mit mir geteilt hat. Mein Freund Omar für all die Laufeinheiten, Rennradtouren und die ansteckende Einstellung, dem Leben mit einer sportlichen Unerbittlichkeit zu begegnen. Mein Cousin Maciek für all die unbeschwerte Zeit, die wir verbracht haben. Von ganzem Herzen möchte ich meinen Eltern für ihre unermüdlichen Bemühungen danken, mich großzuziehen, mich zu unterstützen und mich auf diese Reise vorzubereiten.

Eternal blackness beyond the stars
We think our wisdom will get that far

Iron Maiden - If Eternity Should Fail

Abstract

Although the behavior of agents is often led by self-interest, many environments pose an incentive for cooperation by accomplishing a task together and thus be compensated collectively. This naturally leads to the search of a payout mechanism that assigns to each agent a share of the collective benefit which reflects its individual contribution to the completed task. Game theory models such scenarios by the notion of cooperative games in which the agents are the participating players. Within the game-theoretic framework, the Shapley value poses the most prominent solution to the emerging fair division problem, arguably capturing a widespread understanding of fairness.

Over the last decade, the Shapley value has received unprecedented attention within the field of machine learning, attributing importance to entities such as features, datapoints, and structural components of predictive models. Especially the branch of explainable artificial intelligence picked it up as a means to provide understanding of the decision-making of increasingly complex and opaque models. Likewise, Shapley interactions which capture synergies between players have recently attracted attention. Unfortunately, the computational complexity of both quantities, the Shapley value and Shapley interaction, suffers from the exponential blow-up w.r.t. to the number of involved players and thus becomes quickly infeasible in practice. This incentivizes the research on approximation algorithms that return precise estimates while palpating the cooperative game as little as possible.

In this thesis, we develop approximation algorithms that leverage novel representations of the Shapley value and Shapley interactions on the basis of mean estimation and weighted regression which allow for tailored sampling schemes. Given the Shapley value's richness of applications, our methods are purposefully domain-independent without imposing structural assumptions. Consequently, they can be applied across the entire spectrum of emerging cooperative games. To this end, we place special emphasis on the variance reduction technique of stratification to develop methods that utilize the gathered information from each sample to a richer degree than in other representations possible and derive theoretical guarantees for the estimates' precision. Empirical evaluations in the context of machine learning confirm the soundness of our propositions and their capability to display an advantage over competing methods.

Zusammenfassung

Obwohl das Verhalten von Akteuren oft von Eigeninteresse geleitet ist, setzen viele Szenarien einen Anreiz zur Kooperation, indem Akteure gemeinsam eine Aufgabe bewältigen und dafür kollektiv vergütet werden. Dies führt zwangsläufig zu der Frage nach einem Auszahlungsmechanismus, der jedem Akteur seinen Anteil an der kollektiven Vergütung ausschüttet, welcher dessen individuellen Beitrag zur Bewältigung ebendieser Aufgabe widerspiegelt. Die Spieltheorie modelliert solche Szenarien anhand des Konzepts eines kooperativen Spiels, das die Akteure als die teilnehmenden Spieler umfasst. Innerhalb des spieltheoretischen Rahmens stellt der Shapley-Wert die prominenteste Lösung für das auftretende Problem der gerechten Aufteilung dar, weil dieser ein weit verbreitetes Verständnis von Fairness erfasst.

Der Shapley-Wert hat über das letzte Jahrzehnt hinweg beispiellose Aufmerksamkeit im Bereich des maschinellen Lernens erhalten und wird unter anderem benutzt, um die Wichtigkeit von einzelnen Attributen, Datenpunkten oder sogar strukturellen Komponenten von Prädiktionsmodellen zu messen. Insbesondere das Feld der erklärbaren künstlichen Intelligenz hat diesen als Werkzeug aufgegriffen, um ein Verständnis für die Entscheidungsfindung immer komplexer und undurchsichtiger werdender Modelle zu vermitteln. Ebenso haben Shapley-Interaktionen, welche Synergien zwischen Spielern quantifizieren, an Interesse gewonnen. Bedauerlicherweise leidet die Rechenkomplexität beider Größen unter einer exponentiellen Zunahme in Bezug auf die Anzahl der beteiligten Spieler und wird somit schnell impraktikabel. Dieser Umstand motiviert die Erforschung von Approximationsalgorithmen, die den Shapley-Wert und Shapley-Interaktionen möglichst präzise schätzen.

Diese Arbeit entwickelt Approximationsalgorithmen, die neuartige Darstellungen des Shapley-Wertes und der Shapley-Interaktionen anhand von Mittelwertschätzung und gewichteter Regression nutzen, welche dementsprechend angepasste Stichprobeverfahren zum Schätzen ermöglichen. Angesichts der Vielfalt an Anwendungen des Shapley-Wertes treffen wir keine strukturellen Annahmen über das kooperative Spiel, sodass die entwickelten Methoden bewusst domänenunabhängig und über das gesamte Spektrum an kooperativen Spielen anwendbar sind. Tiefergehend behandeln wir Stratifizierung als Technik zur Varianzreduktion von Schätzern, um Algorithmen zu entwickeln, welche die in den gesammelten Stichproben enthaltene Information zu einem höheren Grad nutzen, als andere Darstellungen dies ermöglichen, und theoretische Garantien für die Approximationsgüte zu geben. Empirische Untersuchungen im Kontext des maschinellen Lernens bestätigen die Fundiertheit unserer Methoden und deren Fähigkeit, konkurrierende Ansätze zu schlagen.

Contents

1 Introduction

Competitive environments seemingly promote selfishness. However, collaboration between agents or parties poses a fruitful business model in many economic scenarios. Smaller logistic service providers, for example, join forces to cost-effectively solve routing problems emerging from transportation demands. In doing so, the involved parties increase their profitability and overcome their cost disadvantage against larger companies (Schopka and Kopfer, 2015; Kimms and Kozeletskyi, 2016). Software firms form joint ventures to benefit from each other's expertise and knowledge exchange while sharing fixed costs (Fahimullah et al., 2019). Electricity providers in energy grids, as another example, cooperate by collectively responding to electricity demands (Bremer and Sonnenschein, 2013; O'Brien et al., 2015), ensuring the functioning of civil infrastructure. A central question is how to distribute the gained profit among agents such that each receives a fair share which reflects its contribution to the collective benefit achieved by the group. Conversely, the presence of an equitable payout mechanism in the first place may incentivize cooperation.

The arising fair division problem is subject of extensive research within the field of game theory. Cooperative games, one of the field's most popular concepts, facilitate axiomatically guided approaches. Within this notion, the collaborating agents are interpreted as abstract players which can form arbitrary subgroups, so-called coalitions, that capture cooperation between the included players. In addition, a value function that assigns a real-valued worth to each possible coalition models the collective benefit a group of players achieves by solving a certain task. Combined with the sheer versatility of this formalism, constructing this value function appropriately enables to model a wide range of fair division problems crossing multiple domains beyond profit allocation, including finance (Moehle et al., 2022) and social networks (Gaskó et al., 2023). The induced lattice of coalition values gives room to impose axiomatic desiderata on an equitable allocation that divides the collective benefit achieved by all players. The Shapley value (Shapley, 1953) emerged as the most prominent division rule, as it is the unique solution to fulfill a certain set of axioms that arguably capture an intuitive understanding of fairness.

Recognizing the Shapley value's axiomatic derivation, the field of machine learning started to employ it for the purpose of constructing explanations that shed light on the intricate inner workings of machine learning models (Sundararajan and Najmi, 2020). Facing the rapid growth in complexity of modern models which have consequently become increasingly opaque, the branch of explainability offers various methods to aid understanding of their decision-making (Vilone and Longo, 2021), thus reclaiming a certain sense of trustworthiness. Among those, additive feature explanations decompose an observed effect such as the predicted value or a model's generalization performance among the features of the data and assign to this end an importance score to each feature. Performing this attribution via the Shapley value has attracted significant interest, leading to a diverse variety of Shapley-based explanations (Rozemberczki et al., 2022). Their appeal stems from the simplicity by which the trade-off between fidelity and readability of an explanation is tackled. Although the simplification to individual importance scores is interpretable to the human user, it compresses the constructed cooperative game behind the fair division problem quite drastically, potentially hiding synergies between features. Fittingly, Shapley interactions (Grabisch and Roubens, 1999) conceptually extend the Shapley value and render the interplay between players tangible, enriching explanations by additionally assigning scores to pairs and triples of features (Fumagalli et al., 2023).

The axiomatic uniqueness of the Shapley value comes with a price to pay in complexity. Its inherent deficiency is rooted in the blow-up of the number of feasible coalitions which scales exponentially with the number of players in the game. In fact, in the absence of drastic restrictions to the value function, the computation of the Shapley value is NP-hard (Deng and Papadimitriou, 1994). As a practical consequence, its applicability is severely limited, if not vanished for large player numbers, as often encountered in datasets of high dimensionality. The dooming infeasibility poses a pressing need to reliably estimate Shapley values and interactions. To this end, approximation algorithms palpate the value function primarily through statistical sampling of coalition values.

Given the richness of applications of the Shapley value and interactions, this thesis contributes approximation algorithms that are not only model-agnostic for explanations in machine learning but also domain-independent. The developed methods are applicable to arbitrary cooperative games regardless of their origin. In this spirit, we restrain from imposing heuristics in the shape of structural assumptions on the value function, as they do not only limit applicability, but also impede the validity of theoretical guarantees which we aspire to provide. Instead, we discover novel presentations of the Shapley value to which we develop tailored sampling

schemes that make more effective use of costly observed samples. Further, we are interested in universal theoretical guarantees on the approximation quality that hold true for any game, hinting at how structural properties of a cooperative game ultimately impact the estimates' precision. In particular, we specialize on the variance reduction technique of stratification, refine it in the context of Shapley values, and demonstrate its hypothetical potential. Our algorithms empirically converge to this optimum and compare favorably to other proposed methods depending on the game's domain. Last but not least, we distinguish between approximating all players' Shapley values precisely and identifying the most influential players according to their Shapley values. The subtle but significant shift in the objective gives room to transfer algorithmic approaches from online learning that we take advantage of.

1.1 Thesis Structure and Contained Works

As this thesis aspires to provide universal approximation methods and investigate their properties detached from domain-specific applications, we start by giving a brief introduction to cooperative game theory in Chapter 2 at a more conceptual level. Being equipped with formal concepts and the axiomatic derivation of the Shapley value and Shapley interaction, this part will hint at alternatives to both quantities and present our considered notion of their ubiquitous approximation problem.

We continue to present commonly appearing constructions of cooperative games within the field of machine learning in Chapter 3. We will touch upon additive feature explanations more thoroughly since these form the major motivation of the Shapley value in machine learning. However, our goal is not to give a comprehensive overview of explainable AI, more so, we want to raise awareness for the intricate differences in modeling cooperative games that substantially impact the interpretation of the resulting explanations. Moreover, we come across other games that not necessarily fulfill an explanatory purpose but are used to perform selection of entities such as features, datapoints, and model components. Both types of fair division problem are part of our empirical evaluations.

Chapter 4 categorizes common approximation methods of the Shapley value. Their categorization serves as a platform for embedding the contribution of this thesis into the context of current state-of-the-art methods. The main contribution of this thesis comprises the following works that are given from Chapter 5 to 11 whose appendices are contained from Appendix A to D:

(I) 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, February 20-27, Vancouver, Canada,* AAAI Press, 2024, pp. 13246–13255.

(II) Patrick Kolpaczki, Georg Haselbeck, and Eyke Hüllermeier. "How Much Can Stratification Improve the Approximation of Shapley Values?". In: *Proceedings of the 2nd World Conference on Explainable Artificial Intelligence, Part II, July 17-19, Valletta, Malta,* Communications in Computer and Information Science. Springer, 2024, pp. 489–512.

(III) Patrick Kolpaczki. "Comparing Shapley Value Approximation Methods for Unsupervised Feature Importance", In: *Proceedings of DataNinja sAIOnARA 2024 Conference, 25-27 June, Bielefeld, Germany,* BieColl – Bielefeld eCollections, 2024, pp. 13-15.

(IV) Guilherme Dean Pelegrina, Patrick Kolpaczki, Eyke Hüllermeier. "Shapley Value Approximation Based on k-Additive Games". In: *CoRR* abs/2502.04763, 2025.

(V) Patrick Kolpaczki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. "SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification". In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistic, AISTATS, 2-4 May, Valencia, Spain,* Volume 238. Proceedings of Machine Learning Research. PMLR, 2024, pp. 3520–3528.

(VI) Patrick Kolpaczki, Viktor Bengs, Eyke Hüllermeier. "Identifying Top-k Players in Cooperative Games via Shapley Bandits". In: *Proceedings of the LWDA 2021 Workshops: FGWM, KDML, FGWI-BIA, and FGIR, September 1-3, Online,* Volume 2993. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 133–144.

(VII) Patrick Kolpaczki, Tim Nielen, Eyke Hüllermeier. "Antithetic Sampling for Top-k Shapley Identification". In: *CoRR* abs/2504.02019, 2025.

At last, we conclude this thesis in Chapter 12 by recapitulating the key findings and advances offered to the field of approximating the Shapley value and Shapley interactions. Directing our attention at future work to succeed this thesis, we point out promising research avenues to further refine the proposed methods.

2 Introduction to Cooperative Game Theory

The notion of a cooperative game forms the foundation of this thesis and the considered problem of approximating the Shapley value and interactions. After providing a glimpse on the fundamentals of cooperative game theory in Section 2.1, we introduce in Section 2.2 the Shapley value as a member of the so-called class of semivalues, an axiomatic class of solution concepts to the fair division problem. Section 2.3 expands further to Shapley interactions by investigating on how to quantify not only the contribution of single players but the interplay of whole groups of players. Finally, Section 2.4 touches upon the computational complexity of the presented game-theoretic quantities and highlights subtle but decisive differences between the task of approximating all Shapley values precisely and just identifying the players with the highest Shapley values.

2.1 Cooperative Games

The frequently appearing situation of agents making agreements in order to accomplish a task with each of them aiming to reap a selfish benefit calls for a systematic approach to model such scenarios of cooperation. Since these agents do not necessarily have to be of human form, as for example robots, or at least not represent individual human beings such as companies, organizations, and even whole states, we will refer to these agents as *players* in a cooperative game. Typically, these players are given by a *player set* $\mathcal{N}$ with each element $i \in \mathcal{N}$ being a player. Within this notion, cooperation comes into existence through the formation of *coalitions* of players that are represented by subsets of $\mathcal{N}$.

A critical but often needed assumption is that each possible coalition $S \subseteq \mathcal{N}$ can not only be formed, but also that its *collective benefit* that S would achieve by (partially) completing the task at hand is measurable. The collective benefit of a coalition can be seen as a joint payout for cooperation and is often called *worth* or *value* of S.

Since each coalition has its own worth, it comes naturally to think of a mapping ν that assigns to each coalition $S \subseteq \mathcal{N}$ its worth $\nu(S)$. The set function ν is commonly referred to as the *characteristic function* or *value function*.

The next assumption that we impose is the transferability of worth between players. The joint payout can be divided and distributed arbitrarily among the players in a coalition. Transferability is often met in practice, as it is common to pay in monetary units for performed work or service. In contrast, this excludes scenarios in which payouts take the shape of indivisible goods. To embed this formally, we assume each worth $\nu(S)$ to be a real-valued number, thus leading to a value function that maps from the power set of players to the set of real numbers, i.e. $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$. Equipped with player set and value function, we can define what is known as a *cooperative game*, *coalitional game*, or *transferable utility game* (Brânzei et al., 2008).

Definition 2.1. Cooperative Game
A cooperative game is given by the pair $(\mathcal{N}, \nu)$ with $\mathcal{N}$ being its set of players and $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ its value function.

The formalism of a cooperative game is simple yet expressive enough to model many scenarios of cooperation due to its abstract nature. To keep this spirit, we denote the player set as $\mathcal{N} = \{1, \dots, n\}$ with n expressing the finite number of players. Hence, a game allows 2^n many coalitions to form. This exponential growth will turn out to be crucial in Section 2.4. The power set contains two special subsets: the *empty coalition* $\emptyset$ and the *grand coalition* $\mathcal{N}$ being the player set itself. The worth $\nu(\emptyset)$ can be challenging to interpret and is often set to zero, capturing the common sense that no work performed should lead to no value. However, we will encounter in Chapter 3 the construction of games that do exhibit a non-zero worth of the empty coalition without running into difficulties of interpretation. On the other side, ν can be scaled such that the worth $\nu(\mathcal{N})$ of the grand coalition is 1, thus essentially becoming a capacity or fuzzy measure (Sugeno, 1974). Note that for any ν with $\nu(\mathcal{N}) - \nu(\emptyset) \neq 0$ both requirements come without loss of generalization since ν can be normalized into ν' as follows:

$$\nu'(S) = \frac{\nu(S) - \nu(\emptyset)}{\nu(\mathcal{N}) - \nu(\emptyset)} \quad \text{for all } S \subseteq \mathcal{N}\,. \tag{2.1}$$

Interestingly and simultaneously important, Definition 2.1 and also the normalization do not restrict ν to be negative. In fact, it is even appropriate to assign negative values to coalitions with a destructive impact, potentially arising through the incompatibility of certain players or the presence of those with malicious intentions.

In order to illustrate cooperative games and the upcoming notions in the remainder of this chapter, we will continuously make use of the exemplary game given by Example 2.2 that could exist in real life in this way or another.

Example 2.2. Three car mechanics decide to start a business with their own automotive workshop. Each of them brings varying levels of experience within the required skillset ranging from performing the actual repairs to business administration. The three of them try to help each other and contribute their part. The player set is thus modeled as $\mathcal{N} = \{1, 2, 3\}$. The worth of a coalition is measured by the monthly revenue the workshop would achieve by being run solely with the mechanics contained within that coalition. The value function is given by Table 2.1.

S	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$\nu(S)$	0	20	40	50	60	80	100	120

Table 2.1.: Tabular representation of the value function for three players.

So far, we have only used cooperative games as a descriptive tool for mathematical modeling. In the next section we will derive quantities that provide further insight into the weighted exponentially-sized lattice of coalitions, illustrated by Figure 2.1, that a game $(\mathcal{N}, \nu)$ spans. Moreover, it is not unusual to impose structural assumptions on the shape of ν (see (Valásková and Struk, 2005) for an overview), for example, one of the more popular being monotonicity. For a *monotone* game holds

$$\nu(S) \leq \nu(T) \quad \text{for all } S, T \subseteq \mathcal{N} \text{ with } S \subseteq T\,. \tag{2.2}$$

However, for the remainder of this chapter, we will consider games with arbitrary value functions that are not restricted to any particular shape.

2.2 The Shapley Value: A Unique Solution

In light of a cooperative game being present with players that seek their own individual benefit, one is predominately faced with the following two questions:

1. Which coalitions are formed by players who agree to cooperate?

2. How is the worth of a realized coalition distributed among its players?

Although extensive studies have been conducted to answer the first question (Ray, 2007), we will consider the grand coalition as given. In other words: all players agree to cooperate together and bring the grand coalition to life. This leaves us with the *fair division problem* of how to divide the gain in collective benefit $\nu(\mathcal{N}) - \nu(\emptyset)$ among all players in $\mathcal{N}$. For the sake of convenience, we will assume $\nu(\emptyset) = 0$ such that simply $\nu(\mathcal{N})$ has to be distributed. Worth mentioning is that this question can not only bee seen from the perspective of distributing worth in equitable manner. One may interpret the assigned value to a player as a measure of his contribution to the fulfillment of a task or even power, for instance within a voting system.

We approach the fair division problem axiomatically and derive possible solutions systematically based on desired properties, partly recapitulating the work of Shapley (1953). We begin with the notion of a *payoff distribution* on which we later impose desiderata.

Definition 2.3. Payoff Distribution
Given a cooperative game $(\mathcal{N}, \nu)$, a payoff distribution $x \in \mathbb{R}^n$ is a real-valued sequence of length n.

Given a specific game, one may search for a suitable payoff distribution that assigns a reward to each player while fulfilling desirable properties. But to be more precise, what we are actually seeking is a mechanism that describes how to form a payoff distribution for any cooperative game we might be confronted with. Hence, we define *solution concepts* which induce a mapping from the space of cooperative games to that of payoff distributions.

Definition 2.4. Solution Concept
A solution concept is a function Γ that maps any cooperative game $(\mathcal{N}, \nu)$ and contained player $i \in \mathcal{N}$ to a real-valued payoff, i.e.

$$\Gamma(\mathcal{N}, \nu, i) \mapsto x_i \in \mathbb{R}\,.$$

We denote the payoff assigned to player i as $\Gamma_i(\mathcal{N}, \nu) := \Gamma(\mathcal{N}, \nu, i)$ or x_i if there is no ambiguity about the considered game. The payoff distribution induced by Γ for a game is thus $\Gamma(\mathcal{N}, \nu) := (\Gamma_1(\mathcal{N}, \nu), \ldots, \Gamma_n(\mathcal{N}, \nu))$ also denoted as x. A trivial example of a solution concept is one that assigns the uniform distribution to all cooperative games. Ergo, each player i is assigned the same payoff $x_i = \frac{\nu(\mathcal{N})}{n}$.

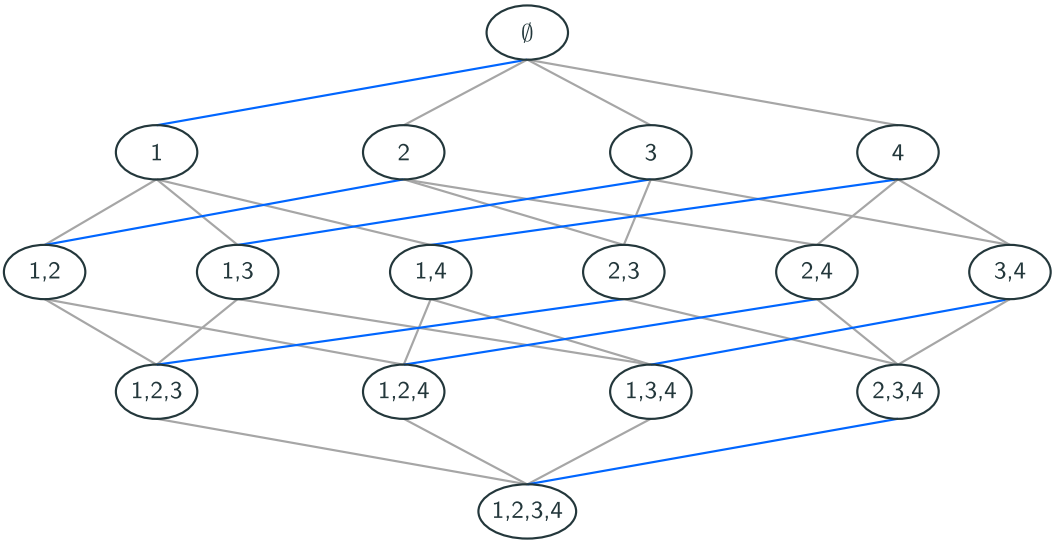


Figure 2.1.: A cooperative game $(\mathcal{N}, \nu)$ spans a lattice of exponential size w.r.t. the number of players n, illustrated here for four players $\mathcal{N} = \{1, 2, 3, 4\}$. Each coalition $S \subseteq \mathcal{N}$ is represented by a node which can be associated with a weight given by its worth $\nu(S)$. Each marginal contribution of a player i to a coalition S forms an edge weighted with $\Delta_i(S)$. The coalitions are grouped by cardinality in layers and the marginal contributions of player 1 are marked in blue.

Unfortunately, this does not reflect that players might contribute to the task at hand to varying degrees. As a consequence, some players may even view this as unfair treatment and be incentivized to complain as demonstrated by Example 2.5.

Example 2.5. Considering the scenario of Example 2.2, the uniform distribution would assign to each of the three car mechanics a payoff of $x_i = 40$, despite the increase in worth that the substitution of player 1 by player 2 causes for each coalition containing player 1 but not player 2. This indicates that player 2 contributes more and thus should also receive a higher payoff.

This brings us to the notion of *marginal contributions* that will allow us to measure the impact of individual players and formalize an intuitive understanding of fairness.

Definition 2.6. Marginal Contribution
Given a cooperative game $(\mathcal{N}, \nu)$, the marginal contribution of a player $i \in \mathcal{N}$ to a coalition $S \subseteq \mathcal{N} \setminus \{i\}$ is given by

$$\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S) \,.$$

Practically speaking, a player's marginal contribution captures the increase in collective benefit caused by him joining that coalition. To preserve a certain meaningfulness, we exclude coalitions which already contain that player, i.e. $i \in S$, since the marginal contribution would automatically turn zero. It can also turn negative if a player causes a loss of collective benefit. As indicated, we can now construct axioms (Shapley, 1953) that arguably capture desiderata of a fair solution concept.

We call $i \in \mathcal{N}$ a *dummy player* if all its marginal contributions are equal, i.e. there exists some $c_i \in \mathbb{R}$ with $\Delta_i(S) = c_i$ for all $S \subseteq \mathcal{N} \setminus \{i\}$, and c_i its *dummy contribution*.

Definition 2.7. Dummy Player Axiom
A solution concept Γ fulfills the dummy player axiom if for any cooperative game $(\mathcal{N}, \nu)$ and player $i \in \mathcal{N}$ it assigns the dummy contribution c_i of i as its payoff, i.e.

$$x_i = c_i \text{ for each dummy player } i \in \mathcal{N}\,.$$

The dummy player axiom enforces the expectation that a player who does not interact with other players and independently contributes its constant part to all coalitions, should not receive more or less than that contribution. One might consider this axiom as mild and lenient since it rarely comes into force. A single deviating marginal contribution of player i suffices to no longer impose any restriction on x_i.

We call two distinct players $i, j \in \mathcal{N}$ *symmetric* if their marginal contributions are equal for each coalition that does not contain both, i.e. $\Delta_i(S) = \Delta_j(S)$ for all $S \subseteq \mathcal{N} \setminus \{i, j\}$.

Definition 2.8. Symmetry Axiom
A solution concept Γ fulfills the symmetry player axiom if for any cooperative game $(\mathcal{N}, \nu)$ and symmetric players $i, j \in \mathcal{N}$ it assigns equal payoff to i and j, i.e.

$$x_i = x_j \text{ for each pair of symmetric players } i, j \in \mathcal{N}\,.$$

Effectively, the symmetry axiom implies that two players, who can be mutually substituted within all coalitions without a change in worth, are granted the same payoff since measured by their marginal contributions, they contribute equally. This axiom excludes solution concepts that are discriminative or biased against certain players, capturing an idea that is at the core of fairness.

Definition 2.9. Linearity Axiom
A solution concept Γ fulfills the linearity axiom if for any player set $\mathcal{N}$, player $i \in \mathcal{N}$, two value functions ν_1, ν_2 for $\mathcal{N}$, and $c \in \mathbb{R}$, scaling ν_1 by c scales the payoff of i by c and the sum of the payoffs assigned to i for the games $(\mathcal{N}, \nu_1)$ and $(\mathcal{N}, \nu_2)$ equals the payoff for the sum of both games, i.e.

$$\Gamma_i(\mathcal{N}, c\nu_1) = c\Gamma_i(\mathcal{N}, \nu_1) \text{ and } \Gamma_i(\mathcal{N}, \nu_1) + \Gamma_i(\mathcal{N}, \nu_2) = \Gamma_i(\mathcal{N}, \nu_1 + \nu_2)\,.$$

The addition of two value functions is to be carried out pointwise, i.e. $(\nu_1 + \nu_2)(S) = \nu_1(S) + \nu_2(S)$ for all $S \subseteq \mathcal{N}$. The linearity axiom, is in contrast to the previous two desiderata the first axiom that cannot be applied on the payoff distribution itself for a particular game but requires the introduction of a solution concept. The linear decomposition ensures that the calculation of the payoff distribution can be separated into independent subgames that do not influence each other.

Despite each axiom being relatively loose when considered individually, the combination of all three is in contrast restrictive enough to give rise to a narrow class of solution concepts known as *semivalues* (Dubey et al., 1981) given in Definition 2.10 which is the only family of solution concepts to fulfill these axioms simultaneously.

Definition 2.10. Semivalue (Dubey et al., 1981)
Given any cooperative game $(\mathcal{N}, \nu)$ and weights $w = (w_0, \ldots, w_{n-1}) \in \mathbb{R}^n$, a solution concept Γ is called a semivalue if it assigns to each player $i \in \mathcal{N}$ the payoff

$$\Gamma_i(\mathcal{N}, \nu) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} w_{|S|} \cdot \Delta_i(S)\,.$$

Note that marginal contributions to coalitions of the same size are required to be weighted equally, as otherwise the symmetry axiom would be violated. In addition to the appeal in connection with the fairness axioms, the semivalues yield a convenient interpretation. The payoff for a player given by any semivalue is a weighted average of the player's marginal contributions. Since any choice of weights w is feasible, there are plenty of solution concepts to choose from, and thus the class of semivalues does not yet provide a satisfying answer to the fair division problem. The most straightforward choice would be to assign uniform weights, leading to the *Banzhaf value*, a prominent representative of the class of semivalues.

Definition 2.11. Banzhaf Value (Banzhaf, 1965)
The Banzhaf value is the solution concept Γ that assigns to any cooperative game $(\mathcal{N}, \nu)$ the payoff distribution $\Gamma(\mathcal{N}, \nu) = \beta$ with the payoff of each $i \in \mathcal{N}$ given by

$$\beta_i := \frac{1}{2^{n-1}} \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \Delta_i(S) \,.$$

For the remainder, we will use the name of any specific semivalue like the Banzhaf value interchangeably for the solution concept Γ, the induced payoff distribution β, and the payoff β_i of a player. Although the simplicity of the Banzhaf value might convince, Example 2.13 uncovers its significant deficiency regarding an obvious axiom that we have left out so far, namely the efficiency axiom.

Definition 2.12. Efficiency Axiom
A solution concept Γ fulfills the efficiency axiom if for any cooperative game $(\mathcal{N}, \nu)$ the assigned payoff distribution summed up over all players equals the worth of the grand coalition, i.e.

$$\sum_{i=1}^{n} \Gamma_i(\mathcal{N}, \nu) = \nu(\mathcal{N}) \,.$$

This definition is specifically tailored for our assumption of $\nu(\emptyset) = 0$. In the general case, efficiency requires the payoffs of all players to sum up to $\nu(\mathcal{N}) - \nu(\emptyset)$.

Example 2.13. Applying the Banzhaf value to the cooperative game of the three car mechanics in Example 2.2, the first mechanic is assigned a payoff of $\beta_1 = 22.5$, the second $\beta_2 = 42.5$, and the third $\beta_3 = 57.5$. The payoff distribution sums up to 122.5, surpassing the revenue that all three of them achieve together. Hence, their automotive workshop would need to take out a loan to pay its three employees.

Essentially, the Banzhaf value is not guaranteed to sum up to the worth of the grand coalition. A simple idea to fix this issue and satisfy the efficiency axiom is to rescale the payoffs by a factor of $\frac{\nu(\mathcal{N})}{\sum_{i=1}^{n} \beta_i}$. Unfortunately, this comes at the price of losing linearity. However, there is a way of constructing a semivalue that incorporates efficiency. It can be achieved by adjusting the weights which yields the arguably most prevalent semivalue, the *Shapley value*.

Definition 2.14. Shapley Value (Shapley, 1953)
The Shapley value is the solution concept Γ that assigns to any cooperative game $(\mathcal{N}, \nu)$ the payoff distribution $\Gamma(\mathcal{N}, \nu) = \phi$ with the payoff of each $i \in \mathcal{N}$ given by

$$\phi_i := \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot \Delta(S) \,.$$

Like all other semivalues, the Shapley value is a weighted average of a player's marginal contributions. Its unprecedented popularity emerges not only from its compliance with the four axioms, but also the fact that it is the *unique* solution concept to fulfill them (Shapley, 1953). No other solution concept satisfies all four axioms.

Theorem 2.15. *(Shapley, 1953)*
The solution concept $\Gamma(\mathcal{N}, \nu) = \phi$ is the only solution concept to simultaneously fulfill the dummy player, symmetry, linearity, and efficiency axiom.

Not only does Theorem 2.15 speak in favor of the Shapley value, it also excludes all other weights w for semivalues if one intransigently insists on efficiency. Yet, in comparison to the Banzhaf value, the utilized weights within the Shapley value might appear at first arbitrary and incomprehensible. One way to shed light on their shape is to observe the mass of weight that is assigned to each coalition size. The sum of weights connected to each coalition size from zero to $n-1$ is exactly $\frac{1}{n}$. Hence, the mass of weight is first uniformly distributed over the sizes, and then uniformly distributed over the marginal contributions to coalitions of that particular size (see Equation 4.4). The approximation techniques based on stratification presented in Section 4.2 take advantage of this particular observation.

Another approach is to derive the weights from a probabilistic perspective. Since the weights sum up exactly to 1, one can interpret them also as a probability distribution for each player. This distribution over marginal contributions can be rediscovered by averaging a player's marginal contributions that appear throughout all permutations of $\mathcal{N}$. The marginal contribution $\Delta_i(\pi)$ of a player i w.r.t. to some permutation $\pi : \mathcal{N} \to \mathcal{N}$, mapping each position j to a player $\pi(j)$, is the increase in worth when i enters the coalition $\text{pre}_\pi(i) := \bigcup_{j=1}^{\pi^{-1}(i)-1} \{\pi(j)\}$ that precedes i in π, hence $\Delta_i(\pi) := \Delta_i(\text{pre}_\pi(i))$. Within this construction each $\Delta_i(S)$ appears potentially multiple times because it can arise from numerous permutations π that cause $\text{pre}_\pi(i) = S$. To be exact, the coalition S can precede i in any arbitrary order, allowing for $|S|!$ different combinations and the irrelevance of the order of

the succeeding players $\mathcal{N} \setminus (S \cup \{i\})$ in π multiplies this number by $(n - |S| - 1)!$. Note that the position of i in π is fixed by the cardinality of S, i.e. $\pi(|S| + 1) = i$. Hence, there exist $|S|! \cdot (n - |S| - 1)!$ many permutations in which each marginal contribution $\Delta_i(S)$ appears. The proportion to the total number of permutations $n!$ leads us back to the weights of the Shapley value, i.e. $\frac{|S|!\cdot(n-|S|-1)!}{n!} = \frac{1}{n\cdot\binom{n-1}{|S|}}$. Consequently, we arrive at a different representation of the Shapley value based on permutations:

$$\phi_i = \frac{1}{n!} \sum_{\pi} \Delta_i(\text{pre}_\pi(i)) \,. \tag{2.3}$$

Hence, the Shapley value in Definition 2.14 can be interpreted as the expected marginal contribution where the randomness is w.r.t. the drawn coalition S not containing i with the probability distribution given by the weights, or w.r.t. permutations of the player set drawn uniformly at random. We demonstrate the latter representation for the calculation of the Shapley value in Example 2.16.

Example 2.16. The Shapley values of the three mechanics in Example 2.2 are $\phi_1 = 21.\bar{6}$, $\phi_2 = 41.\bar{6}$, $\phi_3 = 56.\bar{6}$, and sum up exactly to the worth that they achieve all together. A calculation of ϕ based on permutations is given in Table 2.2.

	Marginal contribution		
Permutation	1	2	3
1, 2, 3	20	40	60
1, 3, 2	20	40	60
2, 1, 3	20	40	60
2, 3, 1	20	40	60
3, 1, 2	30	40	50
3, 2, 1	20	50	50
Average: ϕ_i	$21.\bar{6}$	$41.\bar{6}$	$56.\bar{6}$

Table 2.2.: Tabular calculation of the Shapley value for three players. Each player has its own column with the cell value denoting its marginal contribution when players enter the game in the order of a particular permutation. The Shapley value is the average over all rows, each representing a permutation.

We have systematically answered the question of how to distribute collective benefit among the players of a cooperative game by formalizing desiderata that an aspired fair solution should fulfill. The notion of marginal contributions played a central role, stretching from axioms to the actual solution. We have yet only alluded to the richness of representations for the Shapley value and will encounter further alternatives giving rise to a variety of sampling approaches for approximation in Chapter 4.

2.3 Shapley Interactions: Extension to Higher Order

One would assume that the essence of cooperation lies within the added value that players provide when working together compared to executing partial tasks on their own. It is therefore all the more important to be not mislead by the Shapley value. One could mistakenly interpret each player to simply contribute its Shapley value as the added amount of value in a sequential process of players independently solving a task. The Shapley value might give a convincing solution to the fair division problem but it does not answer how cooperation happens. More precisely we are interested in the synergies between players that arise from cooperation: How do players affect each other in terms of contributed value? We will call these synergy effects *interaction* of players.

The simplest case to consider is the interaction of a pair of players i and j that should be measured by some real number $I_{i,j}$. Starting with pairs, it comes naturally to extend a desired notion of interaction to some higher order, namely to sets of any cardinality beyond pairs. Thus, we carry on in the spirit of Section 2.2 and define with the *interaction index* the equivalent to the solution concept from Definition 2.4.

Definition 2.17. Interaction Index
An interaction index is a function Γ that maps any cooperative game $(\mathcal{N}, \nu)$ and coalition $K \subseteq \mathcal{N}$ to a real-valued interaction I_K, i.e.

$$\Gamma(\mathcal{N}, \nu, K) \mapsto I_K \in \mathbb{R}\,.$$

We will write $\Gamma_K(\mathcal{N}, \nu) := \Gamma(\mathcal{N}, \nu, K)$ which already reveals our intention to extend the Shapley value to interaction since the interaction index subsumes the notion of the solution concept in the case of K being a singleton. Hence, we will also call the interaction $I_i := I_{\{i\}}$ of $\{i\}$ its payoff to be aligned with previous notions.

In contrast to the Shapley value, when speaking about interaction, we are missing the confrontation with a precise problem statement as for fair division. To what question exactly should interaction give an answer to? We will approach this void by formulating our expectations of a suitable interaction index at the example of pairs, making our way to a proposal for sets of any order guided by the rationale behind the concepts in Section 2.2. In doing so, we present the ground-laying work

of Grabisch and Roubens (1999) who establish an axiomatic characterization, albeit following a different structure for didactic reasons.

Taking inspiration from the notion of a player's marginal contribution $\Delta_i(S)$, we first desire to quantify the isolated interaction of a pair $\{i, j\} \subseteq \mathcal{N}$ in the presence of some coalition $S \subseteq \mathcal{N} \setminus \{i, j\}$ by an interaction term $\Delta_{i,j}(S)$. We would surely claim with conviction that i and j do not exhibit any interaction if both players contribute in additive fashion to the worth of S. Hence, we would demand $\Delta_{i,j}(S)$ to be zero if we observe

$$\nu(S) + \Delta_i(S) + \Delta_j(S) = \nu(S \cup \{i, j\})\,. \tag{2.4}$$

Grabisch and Roubens (1999) distinguish two cases should the equality not hold. If the right-hand side is greater, then one can intuitively speak of *profitable cooperation* between i and j since their mutual presence increases the attained worth surpassing the combination of individual contributions. Otherwise, one calls it *harmful cooperation* because the presence of one player impedes the other player to put its contribution to full display. Splitting the marginal contributions into coalition values, we can rearrange Equation 2.4 to

$$\nu(S \cup \{i, j\}) - \nu(S \cup \{i\}) - \nu(S \cup \{j\}) + \nu(S) = 0\,. \tag{2.5}$$

Since the left-hand side equals zero, just as demanded from an interaction term following our intuition, we shall adopt it as our aspired expression of $\Delta_{i,j}(S)$. This is aligned with our case distinction by the switch in sign of $\Delta_{i,j}(S)$ for profitable and harmful cooperation, further exemplified by Example 2.18.

Example 2.18. The three car mechanics of Example 2.2 share different working attitudes, complement each other in their skills, or sometimes exhibit redundant capabilities leading to all three types of pairwise interactions. Mechanics 1 and 2 have negative interaction of $\Delta_{1,2}(\{3\}) = -10$ in the presence of mechanic 3, mechanics 1 and 2 have no interaction on their own, i.e. $\Delta_{1,2}(\emptyset) = 0$, and mechanics 2 and 3 have positive interaction of $\Delta_{2,3}(\emptyset) = 10$ on their own.

Interestingly, the pairwise interaction term can also be stated by only utilizing either marginal contributions of player i or that of j (Kojadinovic, 2003):

$$\Delta_{i,j}(S) = \Delta_i(S \cup \{j\}) - \Delta_i(S) \tag{2.6}$$

$$\text{and} \quad \Delta_{i,j}(S) = \Delta_j(S \cup \{i\}) - \Delta_j(S)\,. \tag{2.7}$$

This observation yields a convenient interpretation. While the marginal contribution measures the increase in worth of S caused by the presence of i, the interaction of i and j measures the increase of j's marginal contribution to S caused by the presence of i and vice versa. Thus, one might view this as a recursive extension to pairs and we shall bring it to completeness for any cardinality, leading us to the *discrete derivative* that generalizes the marginal contribution from Definition 2.6.

Definition 2.19. Discrete Derivative [1]
Given a cooperative game $(\mathcal{N}, \nu)$, the discrete derivative of a coalition $K \subseteq \mathcal{N}$ to a disjoint coalition $S \subseteq \mathcal{N} \setminus K$ is given by $\Delta_K(S) := \nu(S)$ for $K = \emptyset$ and otherwise for non-empty K and any $i \in K$ by

$$\Delta_K(S) := \Delta_{K \setminus \{i\}}(S \cup \{i\}) - \Delta_{K \setminus \{i\}}(S)\,.$$

Hence, the interaction of K can be stated as the impact of any $i \in K$ on the next lower order interaction of $K \setminus \{i\}$, ultimatively ending in the impact on coalition values. Note that this definition indeed entails the marginal contribution as interaction of singletons. In contrast, for sets of higher cardinality one might question the well-definedness since any $i \in K$ can be chosen to enter the next recursive step. We resolve these doubts by offering an equivalent closed-form representation of $\Delta_K(S)$ that simply sums up coalition values with alternating signs.

Proposition 2.20. *(Kojadinovic, 2003)*
For all cooperative games $(\mathcal{N}, \nu)$ and disjoint coalitions $K, S \subseteq \mathcal{N}$, the discrete derivative $\Delta_K(S)$ is equal to

$$\Delta_K(S) = \sum_{W \subseteq K} (-1)^{|K \setminus W|} \cdot \nu(S \cup W)\,.$$

Now that we have established with the discrete derivative the essential building block for interaction as we did with the marginal contribution for payoff, it is the obvious next step to plug it into the semivalue. This approach follows the thought that the interaction index for any K should likewise be a composition of interaction terms over all coalitions $S \subseteq \mathcal{N} \setminus K$ that come into question, generalizing the summation over all $S \subseteq \mathcal{N} \setminus \{i\}$ in Definition 2.10. By doing so, we construct the class of *cardinal interaction indices*. For convenience, let $[a] := \{b \in \mathbb{N} \cup \{0\} : b \leq a\}$.

[1] We deviate from the definition provided by Grabisch and Roubens (1999) to emphasize the recursiveness. However, both are equivalent as shown by Proposition 2.20.

Definition 2.21. Cardinal Interaction Index (Grabisch and Roubens, 1999)
Given any cooperative game $(\mathcal{N}, \nu)$ and weights $w_{k,s} \in \mathbb{R}$ for each $k \in [n], s \in [n-k]$, an interaction index Γ is called a cardinal interaction index if it assigns to each $K \subseteq \mathcal{N}$ the interaction

$$\Gamma_K(\mathcal{N}, \nu) = \sum_{S \subseteq \mathcal{N} \setminus K} w_{|K|,|S|} \cdot \Delta_K(S)\,.$$

This construction might appear admittedly ad-hoc as it transfers the semivalue to interaction without further justification. Fortunately, the class of cardinal interaction indices enjoys a similar axiomatic basis encompassing adaptations of the dummy, symmetry, and linearity axiom. Seeking a generalization of semivalues, they contain the special case of a singleton K and thus extend those given in Section 2.2.

Definition 2.22. Dummy Interaction Axiom
An interaction index Γ fulfills the dummy interaction axiom if for any cooperative game $(\mathcal{N}, \nu)$ and dummy player $i \in \mathcal{N}$ it assigns the dummy contribution c_i of i as its payoff and assigns zero interaction to all sets containing i, i.e.

$$I_i = c_i \text{ and } I_{K \cup \{i\}} = 0 \; \forall K \in \mathcal{P}(\mathcal{N} \setminus \{i\}) \setminus \{\emptyset\} \text{ for each dummy player } i \in \mathcal{N}\,.$$

The newly included property of interaction with dummy players is under no circumstances groundless. If a player i does not interact with the other coalition members K and simply additively contributes its part, then its presence should make no difference to the interaction between the players in K. Fittingly, this is already captured by the discrete derivative, i.e. $\Delta_K(S \cup \{i\}) = \Delta_K(S)$ with dummy player i.

We call two not necessarily disjoint coalitions $K_1, K_2 \in \mathcal{N}$ of the same size *symmetric* if all their discrete derivatives are pairwise equal for each coalition not containing both, i.e. $\Delta_{K_1}(S) = \Delta_{K_2}(S)$ for all $S \subseteq \mathcal{N} \setminus (K_1 \cup K_2)$.

Definition 2.23. Symmetry Interaction Axiom
An interaction index Γ fulfills the symmetry interaction axiom if for any cooperative game $(\mathcal{N}, \nu)$ and symmetric coalitions $K_1, K_2 \subseteq \mathcal{N}$ it assigns equal interaction to K_1 and K_2, i.e.

$$I_{K_1} = I_{K_2} \text{ for each pair of symmetric coalitions } K_1, K_2 \in \mathcal{N}\,.$$

The equivalence in discrete derivatives implies that the two coalitions can be mutually substituted without a change in value. Hence, the game is essentially oblivious to the choice between them and thus both should be assigned the same interaction by a fair interaction index. Lastly, we again consider linearity which comes with no further extension to interaction.

Definition 2.24. Linearity Interaction Axiom
An interaction index Γ fulfills the linearity interaction axiom if for any player set $\mathcal{N}$, coalition $K \in \mathcal{N}$, two value functions ν_1, ν_2 for $\mathcal{N}$, and $c \in \mathbb{R}$, scaling ν_1 by c scales the induced payoffs by c and the sum of interactions assigned to K for the games $(\mathcal{N}, \nu_1)$ and $(\mathcal{N}, \nu_2)$ equals the interaction for the sum of both games, i.e.

$$\Gamma_K(\mathcal{N}, c\nu_1) = c\Gamma_K(\mathcal{N}, \nu_1) \text{ and } \Gamma_K(\mathcal{N}, \nu_1) + \Gamma_K(\mathcal{N}, \nu_2) = \Gamma_K(\mathcal{N}, \nu_1 + \nu_2)\,.$$

Any interaction index that fulfills these three axioms is of the shape of a cardinal interaction index (Grabisch and Roubens, 1999). The class of cardinal interaction indices distinguishes its members only by their weights $w_{k,s}$ leaving a wide selection of indices to measure interaction. Again, a pragmatic choice would be to use uniform weights $w_{k,s} = \frac{1}{2^{n-k}}$ that add to 1 such that we obtain an arithmetic mean of discrete derivatives. The resulting index is known as the *Banzhaf interaction index*.

Definition 2.25. Banzhaf Interaction Index (Grabisch and Roubens, 1999)
The Banzhaf interaction index is the interaction index that assigns to any coalition $K \subseteq \mathcal{N}$ with $|K| = k$ of a cooperative game $(\mathcal{N}, \nu)$ the interaction

$$I_K^\beta := \frac{1}{2^{n-k}} \sum_{S \subseteq \mathcal{N} \setminus K} \Delta_K(S)\,.$$

Note that the weights are uniform w.r.t. the reference coalition S but do depend on the size of K since k determines the number of discrete derivatives for K. Aligned with our paradigm to generalize the Banzhaf and Shapley value to interaction, the Banzhaf interaction index reduces to the former for $k = 1$. However, the derivation of weights to obtain a *Shapley interaction index* is less straightforward. In comparison to the fair division problem, there is no obvious equivalent to the efficiency axiom (see Definition 2.12) for interaction.

A further axiom lifts the recursive rationale of the discrete derivative to the index itself. Although the Banzhaf interaction index still fulfills it, the recursive interaction axiom will prepare to force a unique choice of weights within the class of cardinal interaction indices with the help of the Shapley value.

Definition 2.26. Recursive Interaction Axiom
An interaction index Γ fulfills the recursive interaction axiom if for any cooperative game $(\mathcal{N}, \nu)$, coalition $K \subseteq \mathcal{N}$ with $|K| \geq 2$, and $i \in K$ holds

$$\Gamma_K(\mathcal{N}, \nu) = \Gamma_{K\setminus\{i\}}(\mathcal{N} \setminus \{i\}, \nu^{+i}) - \Gamma_{K\setminus\{i\}}(\mathcal{N} \setminus \{i\}, \nu^{-i})$$

with $\nu^{+i}, \nu^{-i} : \mathcal{P}(\mathcal{N} \setminus \{i\}) \to \mathbb{R}$, $\nu^{+i}(S) = \nu(S \cup \{i\})$, and $\nu^{-i}(S) = \nu(S)$ for all $S \subseteq \mathcal{N} \setminus \{i\}$.

The recursive interaction axiom requires that the interaction of K equals the difference in interaction of $K \setminus \{i\}$ between the two games in which i is always present and, respectively, is always absent. The combination of these four axioms plus the reduction to the Shapley value for singletons results in the Shapley interaction index.

Definition 2.27. Shapley Interaction Index (Grabisch and Roubens, 1999)
The Shapley interaction index is the interaction index that assigns to any coalition $K \subseteq \mathcal{N}$ with $|K| = k$ of a cooperative game $(\mathcal{N}, \nu)$ the interaction

$$I_K^{\phi} := \sum_{S \subseteq \mathcal{N}\setminus K} \frac{1}{(n-k+1) \cdot \binom{n-k}{|S|}} \cdot \Delta_K(S)\,.$$

Both interaction indices exhibit an axiomatic justification. The result in Theorem 2.28 concludes our excursion from interaction effects relying on discrete derivatives, baked into semivalues, to a Shapley-based notion of interaction for any cardinality.

Theorem 2.28. *(Grabisch and Roubens, 1999)*
The interaction indices I^{β} and I^{ϕ} are the only interaction indices to simultaneously fulfill the dummy interaction, symmetry interaction, linearity interaction, recursive interaction axiom, and in the case of singletons reduce to the Banzhaf value and, respectively, to the Shapley value.

Example 2.29. The three car mechanics of Example 2.2 share the following Banzhaf and Shapley interactions given in Table 2.3. The interactions for all strict subsets $K \subset \mathcal{N}$ do not only potentially differ in magnitude, but can also exhibit different signs. Both indices always assign $\Delta_{\mathcal{N}}(\emptyset)$ as the interaction of the grand coalition.

K	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
I_K^β	22.5	42.5	57.5	-5	5	5	-10
I_K^ϕ	$21.\overline{6}$	$41.\overline{6}$	$56.\overline{6}$	-5	5	5	-10

Table 2.3.: Tabular representation of the Banzhaf and Shapley interactions for three players.

One might observe that the weights in Definition 2.27 are closely linked to those of the Shapley value in Definition 2.14 since not only do they coincide for $k = 1$ but also both intuitive descriptions through mass of weights and permutations which we gave in Section 2.2 are likewise applicable. Any coalition K has discrete derivatives to coalitions S ranging from size zero to $n - k$. The mass of weights is distributed uniformly over these $n - k + 1$ many sizes, and for each size s the mass of weights is again uniformly distributed among the $\binom{n-k}{s}$ coalitions sharing the same size s and not including any player of K.

The discrete derivative of some K in a permutation of players can be understood analogously to the marginal contribution. Here, we merge K to a new player $[K]$ such that we consider permutations π of $\mathcal{N} \setminus K \cup \{[K]\}$. Its discrete derivative is given by $\Delta_K(\pi) := \Delta_K(\text{pre}_\pi([K]))$ with the reference coalition $\text{pre}_\pi([K])$ being the players that precede $[K]$ in π. We count again the number of permutations π that exhibit $\Delta_K(\pi) = \Delta_K(S)$ for any fixed $S \subseteq \mathcal{N} \setminus K$ with $|S| = s$. There exist $s!$ many orderings for the preceding players S and $(n - k - s)!$ many orderings for the succeeding ones $\mathcal{N} \setminus (K \cup S)$. Therefore, we have $|s!| \cdot (n - k - |s|)!$ suitable permutations against $(n - k + 1)!$ permutations in total. This ratio is equal to the weight $w_{k,s} = \frac{1}{(n-k+1)\cdot\binom{n-k}{s}}$ such that the Shapley interaction index can be viewed as the expected discrete derivative for randomly sampled permutations:

$$I_K^\phi = \frac{1}{(n-k+1)!} \sum_\pi \Delta_K(\text{pre}_\pi([K])) \,. \tag{2.8}$$

2.4 Computational Complexity: Approximation as a Resort

We have introduced the Shapley value and Shapley interactions as measures to quantify the contribution of a player and the interaction between players based on desiderata capturing an intuitive notion of fairness. Unfortunately, their axiomatic uniqueness comes with a price to pay. As Figure 2.1 illustrates, the lattice spanned by a cooperative game grows exponentially w.r.t. the number of players n. And since all semivalues and cardinal interaction indices (with non-zero weights) include the worth of all feasible coalitions in their definitions, the exponential growth of the power set bears consequences on their computation. In fact, the exact computation of the Shapley value is even NP-hard as shown by Deng and Papadimitriou (1994) and this result can analogously be extended to the Shapley interaction index. The emerging practical implication is the quickly arising infeasibility of the Shapley value and interaction for growing player numbers.

The evident remedy to overcome this burden is the approximation of Shapley-based measures which we consider in this section as a task given to an approximation algorithm. For simplicity, we will cover it only for the Shapley value since the problem can readily be transferred to Shapley interactions. In particular, we present in the following two similar problem statements of which each comes with their own challenges, namely the *approximate-all* and the *top-k identification* problem.

Approximate-all. Starting with the former, in essence, the natural goal of an approximation algorithm $\mathcal{A}$ is to provide precise estimates by only sparsely palpating the cooperative game's lattice with as few evaluated coalition values as possible. We denote by $\hat{\phi}_i$ the algorithm's estimate of player i's Shapley value ϕ_i for a fixed game $(\mathcal{N}, \nu)$ and by $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_n)$ the estimated payoff distribution for all players. Motivated by the idea that $\mathcal{A}$ should observe only a limited part of the game's lattice, we consider the *fixed-budget* setting. It states that the algorithm is provided $\mathcal{N}$ and access to ν but can evaluate the value function to retrieve the worth of a coalition only a limited number of times that is given by the *budget* $T \in \mathbb{N}$. More precisely, it can choose a sequence of coalitions $S_1, \ldots, S_T$, possibly containing duplicates, from which it sequentially obtains the worth $\nu(S_t)$ of each contained coalition S_t. Hence, an approximation algorithm has two degrees of freedom to maximize the precision of $\hat{\phi}$: the choice of the *retrieval sequence* and how it aggregates the obtained values to an estimate. Important to note is that the retrieval sequence is not determined in advance but can be constructed randomly step by step, and even further, the

distribution from which the next coalition is drawn can be adjusted to previously observed coalition values.

A key element in this task is the error measure $\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ that quantifies the imprecision $\rho(\hat{\phi}, \phi)$ of a returned estimate $\hat{\phi}$ w.r.t. the actual Shapley value ϕ. An ubiquitous property of the error measure within the literature is its additive decomposition in terms of the individual errors for the players. However, different proposals have been made for the individual error themselves and their weighting. The by far most common combination is that of the arithmetic mean of the individual squared errors which results in

$$\rho_{\text{avg,sqr}}(\hat{\phi}, \phi) := \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\phi}_i - \phi_i\right)^2 . \tag{2.9}$$

As an alternative Campen et al. (2018) investigate the absolute individual error and optionally put it into relation with the magnitude of the player's Shapley value, thus forming a percentage error that leads to

$$\rho_{\text{per,abs}}(\hat{\phi}, \phi) := \sum_{i=1}^{n} \frac{|\hat{\phi}_i - \phi_i|}{|\phi_i|} . \tag{2.10}$$

The latter captures the idea of punishing the estimates' deviations increasingly for smaller absolute Shapley values because these have a relatively higher impact. On the other side, the former uniform weights treat each player equally important in the error measure.

Under the possibly random nature of an approximation algorithm $\mathcal{A}$, the returned estimate $\hat{\phi}$ effectively becomes a random variable. Hence, $\mathcal{A}$ should not be seen as a deterministic function $\mathcal{A}(\mathcal{N}, \nu, T) \mapsto \hat{\phi} \in \mathbb{R}^n$ and thus the error measure needs an extension to judge the performance of $\mathcal{A}$. This is typically done by taking the expectation of the error, which is of course significantly different from the error of the expected estimate. Applying this to the arithmetic mean of individual squared errors, we arrive at the expected *mean squared error* (MSE)

$$\mathbb{E}[\text{MSE}] := \mathbb{E}\left[\rho_{\text{avg,sqr}}(\hat{\phi}, \phi)\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \tag{2.11}$$

whose minimization constitutes the goal of an approximation algorithm. Its popularity stems from the fact that it is analytically accessible and yields the bias-variance decomposition of the expected individual error:

$$\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] = \underbrace{\left(\mathbb{E}\left[\hat{\phi}_i\right] - \phi_i\right)^2}_{\text{bias}} + \underbrace{\mathbb{V}\left[\hat{\phi}_i\right]}_{\text{variance}} \tag{2.12}$$

Not only does this incentivize the investigation of bias and variance as properties of $\mathcal{A}$, it also facilitates the empirical evaluation of $\mathcal{A}$ for larger unstructured games. At first sight, the MSE requires ϕ to be known in order to calculate an error. However, this can quickly become infeasible for games with a large number of players n in combination with a missing closed-form polynomial solution of ϕ. If $\mathcal{A}$ is known to be unbiased, i.e. the bias in Equation 2.12 turns zero, the error collapses to the estimate's variance. In practice, one can then empirically measure the variance of $\hat{\phi}_i$ through multiple approximation runs without the need to laboriously compute ϕ_i in advance.

The expectation is only one quantity to summarize the error distribution of $\mathcal{A}$. Another approach is to consider the cumulative distribution by stating an upper bound $\delta \in [0, 1]$ on the probability that an error greater than ε occurs. An algorithm satisfying this condition is called a *probably approximate correct* (PAC) learner. Most often, the individual absolute deviation is employed as the error measure, leading to the condition

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| > \varepsilon\right) \leq \delta\,. \tag{2.13}$$

Before turning our attention to the top-k identification problem, we want to raise awareness why it is appropriate to count the number of accesses to ν instead of the time passed during the execution of $\mathcal{A}$. The following reasons speak in favor of the discrete budget T:

- The actual runtime of an executed algorithm might vary strongly depending on the implementation and used optimizations related to programming language and hardware. Therefore, a comparison might mislead and favor carefully crafted implementations instead of conceptually advantageous methods.

- The approximation quality in dependence of the budget is much easier to analyze and allows for the direct usage of concentration inequalities from probability theory by interpreting the construction of the random retrieval sequence as a sampling procedure.

- In many practical applications, the evaluation of a coalition's worth poses the bottleneck regarding resource and time consumption. This is especially the case for the inference of complex machine learning models and even more so for re-training on large datasets being performed during each access. Hence, the operations performed by $\mathcal{A}$ between evaluations become negligible.

- Eventually, the deciding resource of consumption might not be time but measured in monetary units to which the algorithmic operations barely contribute. For instance, users that investigate the behavior of a remotely offered machine learning model need to pay for its inference that is provided as a service.

Top-k identification. At the core of the approximate-all problem is the precision of each real-valued estimate $\hat{\phi}_i$. A somewhat less restrictive goal would be to correctly observe the relation between the players' Shapley values instead of their actual magnitude. For example, given a cooperative game, one could only be interested in the identity of the player possessing the highest Shapley value. Taking this thought further, the task can be generalized to finding the k players with the highest Shapley values. One can motivate the idea by the potential disinterest towards players that exhibit relatively low contribution which grow in number for games whose distribution of Shapley values is skewed with a few players having high impact and the others forming a uniform mass that sinks into oblivion. The problem setting resulting from this paradigm shift is known as the top-k identification problem. We will present in the following necessary notions to describe it and point out differences to the approximation problem introduced above. The object of interest is now the non-empty coalition $\mathcal{K}^* \subseteq \mathcal{N}$ of size $k \leq n$ such that no other player in $\mathcal{N} \setminus \mathcal{K}^*$ possesses a greater Shapley value than any player in $\mathcal{K}^*$. Naively, under the assumption of mutually distinct Shapley values, one would utilize orderings of $\mathcal{N}$ to characterize $\mathcal{K}^*$ as

$$\mathcal{K}^* := \{\pi(1), \ldots, \pi(k) \mid \pi : \mathcal{N} \to \mathcal{N} \text{ with } \phi_{\pi(i)} > \phi_{\pi(j)} \text{ for all } i < j \in \mathcal{N}\}\,. \quad (2.14)$$

Hence, the algorithm's goal is to identify the top-k players correctly by returning a coalition $\hat{\mathcal{K}} = \mathcal{K}^*$. Obviously, the strictly descending ordering of players π is non-existent as soon as a set of players shares the same Shapley value. Such a set is only critical if it crosses the top-k border, more precisely if fewer than k many players with higher value and fewer than $n - k$ with lower value exist. Confronted with this difficulty, we need to reformulate our goal since now multiple correct answers $\mathcal{K}$

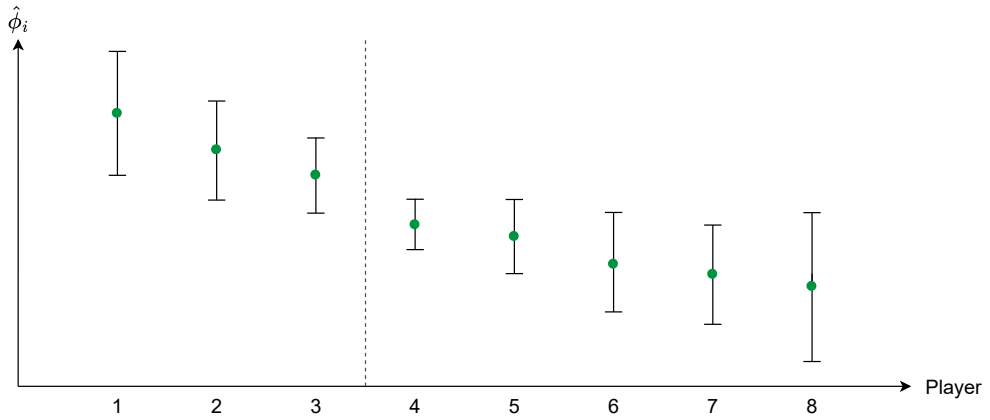


Figure 2.2.: Exemplary illustration of an algorithm's state confronted with the top-k identification problem for $n = 8$ and $k = 3$: The exemplary algorithm $\mathcal{A}$ maintains an estimate $\hat{\phi}_i$ (green dots) and a confidence interval (whiskers) for each player $i \in \mathcal{N}$. The players are sorted in descending order of $\mathcal{A}$'s estimates. As it is the task of $\mathcal{A}$ to separate the three players with the highest Shapley value (dotted line), it can sacrifice the estimate's precision of any player whose confidence interval already strongly indicates to which side it belongs. Here, player 1 is with high confidence within the top-3 because its confidence interval does not intersect with any interval of players to be estimated outside the top-3. Vice versa, player 8 at the bottom end can be likewise excluded from the top-3.

may exist. We abstain from partial orderings and call any coalition $\mathcal{K} \subseteq \mathcal{N}$ of size k *eligible* if it maximizes the sum of contained Shapley values, i.e. it holds

$$\sum_{i \in \mathcal{K}} \phi_i = \max_{S \subseteq \mathcal{N} : |S| = k} \sum_{i \in S} \phi_i \,, \tag{2.15}$$

and denote the set of all eligible coalitions as $\mathcal{E}_k \subseteq \mathcal{P}(\mathcal{N})$. This maximal sum of k many Shapley values

$$\Phi_k := \max_{S \subseteq \mathcal{N} : |S| = k} \sum_{i \in S} \phi_i \tag{2.16}$$

is unique and shared by all $\mathcal{K} \in \mathcal{E}_k$. Note that in case of mutually distinct Shapley values only $\mathcal{K}^*$ is eligible, i.e. $\mathcal{E}_k = \{\mathcal{K}^*\}$. Consequently, the approximation algorithm's task is to return a coalition $\hat{\mathcal{K}} \in \mathcal{E}_k$. We stick to the fixed-budget setting, meaning that the algorithm has again a budget T at its disposal.

What significantly changes, is the error measure whose meaning we invert to obtain a real-valued precision $\psi(\hat{\mathcal{K}}, \mathcal{E}_k)$ that is supposed to express how close the returned

estimate $\hat{\mathcal{K}}$ comes to being eligible and thus to be maximized. The simplest yet most strictest measure to propose is the *binary precision*

$$\psi_{\text{bin}}(\hat{\mathcal{K}}, \mathcal{E}_k) := \begin{cases} 1 & \text{if } \hat{\mathcal{K}} \in \mathcal{E}_k \\ 0 & \text{if } \hat{\mathcal{K}} \notin \mathcal{E}_k \end{cases} \tag{2.17}$$

which harshly punishes every mistakenly included player within $\hat{\mathcal{K}}$. It does not differentiate between non-eligible coalitions of different degree and thus it provides only limited guidance to investigate the quality of differing algorithms. Instead, and to rectify this shortcoming, one can consider the *ratio precision*

$$\psi_{\text{rat}}(\hat{\mathcal{K}}, \mathcal{E}_k) = \max_{\mathcal{K} \in \mathcal{E}_k} \frac{|\hat{\mathcal{K}} \cap \mathcal{K}|}{k}\,, \tag{2.18}$$

stating what percentage of players in $\hat{\mathcal{K}}$ does not need to be swapped out and substituted by players from $\mathcal{N} \setminus \hat{\mathcal{K}}$ to form an eligible coalition. One may view it as a refinement of the prior since it extends the effective codomain of the precision measure from $\{0, 1\}$ to the unit interval $[0, 1]$. Still, the ratio precision misses out on an opportunity to further distinguish the quality of estimates. For example, two non-eligible coalitions $\hat{\mathcal{K}}_1$ and $\hat{\mathcal{K}}_2$ may share the same precision but the falsely included players in $\hat{\mathcal{K}}_1$ possess greater Shapley values than those falsely included in $\hat{\mathcal{K}}_2$. In the spirit of Equation 2.15, $\hat{\mathcal{K}}_1$ comes closer to being eligible because the sum of its Shapley values is greater and thus closer to Φ_k. This blind spot is covered by the *relative precision*

$$\psi_{\text{rel}}(\hat{\mathcal{K}}, \mathcal{E}_k) := \frac{\sum_{i \in \hat{\mathcal{K}}} \phi_i}{\Phi_k}\,, \tag{2.19}$$

defined for all cooperative games with $\Phi_k \neq 0$. This measure can turn negative if the sum for the estimate $\hat{\mathcal{K}}$ is negative but $\Phi_k > 0$. In contrast to the previous measures, the relative precision is not bounded over all games. It can be arbitrarily large for $\Phi_k < 0$ and arbitrarily small for $\Phi_k > 0$. In the former case it even needs to be minimized instead of being maximized. A further downside compared to the ratio precision is its inability to count how many players are correctly identified.

Kariyappa et al. (2024) propose two error measures that instead of counting the wrongly included players, similar to Equation 2.18, indicate the lost contribution in Shapley value. For this purpose, let $\phi_{k^*} := \min_{\mathcal{K} \in \mathcal{E}_k} \min_{i \in \mathcal{K}} \phi_i$ be the smallest Shapley value associated with the players contained in any eligible coalition. Obviously,

it can be found in all eligible coalitions. The *inclusion error* measures the maximal deficit in Shapley value of a player in $\hat{\mathcal{K}}$ compared to ϕ_{k^*}:

$$\rho_{\text{inc}}(\hat{\mathcal{K}}, \mathcal{E}_k) := \inf \{\varepsilon \in \mathbb{R}^{\geq 0} \mid \phi_i \geq \phi_{k^*} - \varepsilon \; \forall i \in \hat{\mathcal{K}}\} . \tag{2.20}$$

On the other hand, the *exclusion error* measures the maximal advantage in Shapley value of a player not in $\hat{\mathcal{K}}$ compared to ϕ_{k^*}:

$$\rho_{\text{exc}}(\hat{\mathcal{K}}, \mathcal{E}_k) := \inf \{\varepsilon \in \mathbb{R}^{\geq 0} \mid \phi_i \leq \phi_{k^*} + \varepsilon \; \forall i \in \mathcal{N} \setminus \hat{\mathcal{K}}\} . \tag{2.21}$$

Kariyappa et al. (2024) combine both measures to form the *inclusion-exclusion error* that takes the maximum of both:

$$\rho_{\text{inc+exc}}(\hat{\mathcal{K}}, \mathcal{E}_k) := \max\{\rho_{\text{inc}}(\hat{\mathcal{K}}, \mathcal{E}_k), \rho_{\text{exc}}(\hat{\mathcal{K}}, \mathcal{E}_k)\} . \tag{2.22}$$

Just as for the approximate-all task, these measures have to be lifted to the randomness of the approximation algorithm $\mathcal{A}$. Again, the expectation and the PAC notion are suitable choices. Here, the expected binary precision yields a convenient interpretation. It specifies the probability of $\mathcal{A}$ correctly returning an eligible coalition. This immediately follows from the observation that the binary precision of $\mathcal{A}$ is effectively a Bernoulli random variable.

A crucial difference between the approximate-all and the task of top-k identification is that the latter does not imply the necessity of precise estimates ϕ_i for the players' Shapley values, at least not for all of them. It suffices to correctly identify the order relations $\phi_i \geq \phi_j$ between the players with the precision of the estimates $\hat{\phi}_i$ playing a minor role. This setting allows to sacrifice the individual estimates' precision for players with Shapley values at the very bottom (or very top) of the spectrum because they can be confidently excluded from the top-k (or assigned to it), and use the saved budget to improve the accuracy for players in the proximity of ϕ_{k^*}. See Figure 2.2 for an illustration of this idea.

In contrast, the approximate-all task requires each individual estimate $\hat{\phi}_i$ to be as precise as possible. Here, the algorithm $\mathcal{A}$ can prioritize players depending on the weighting of the individual errors and the observed sample variance for each estimate. Finally, every approximation algorithm for the approximate-all problem can be adapted for the top-k identification by simply returning the k players with the highest estimates $\hat{\phi}_i$. This transfer might fail vice versa since no algorithm for the latter problem is obliged to maintain an estimate $\hat{\phi}$.

3 Cooperative Games in Machine Learning

The Shapley value, and more so cooperative games in general, are recognized as viable formalisms to model collaboration and define payouts spanning over a wide range of practical scenarios. Classical examples include applications in economics where profit (Bremer and Sonnenschein, 2013; O'Brien et al., 2015; Fahimullah et al., 2019) or cost (Schopka and Kopfer, 2015; Kimms and Kozeletskyi, 2016) is to be fairly allocated among participating parties of joint ventures. Moreover, the Shapley value found its way into finance as a tool to attribute performance or risk to individual assets held in a portfolio (Shalit, 2020; Shalit, 2021; Moehle et al., 2022). Not necessarily seeking an equitable distribution, it is used to detect the most influential individuals within social networks (Campen et al., 2018; Gaskó et al., 2023).

Over the last decade, the Shapley value has sparked significant interest in the field of machine learning. Its rising prominence has been fueled by the discovery that it can be used to construct feature explanations, and thus hopefully shed light onto the decision-making of complex black-box models. We discuss the construction of cooperative games for the purpose of additive feature explanations in Section 3.1 and how interactions enrich these explanations overcoming potential deficiencies of the Shapley value in Section 3.2. Besides providing understanding to the human user, the Shapley value and interactions can also serve a more performance-driven purpose by quantifying the contribution of individual features to a generalization task, thus guiding feature selection. This approach can likewise be applied to entities such as datapoints and components of a model, which we touch upon in Section 3.3.

3.1 Additive Feature Explanations

In the context of the fair division problem, the Shapley value is often applied or interpreted from a *normative* perspective. It determines the share ϕ_i of the collective benefit $\nu(\mathcal{N})$ that each player $i \in \mathcal{N}$ participating in the cooperative game should

receive as an earned payout reflecting its own contribution to the group. However, this is not necessarily the only direction of interpretation. Instead, by departing from the motivation of distributing fair payouts to reward-seeking agents, one can not only ask how much each player should be compensated, but rather try to understand how much each player contributes to $\nu(\mathcal{N})$. Consequently, this view confers a *descriptive* meaning on the Shapley value, transforming it into an instrument to quantitatively investigate how cooperation takes place and how players participate in that.

Practically speaking, this allows to decompose a system's observed behavior measured by some numerical effect $\nu(\mathcal{N})$ among contributing factors $\mathcal{N} = \{1, \dots, n\}$ in an additive manner. Within machine learning, the branch of explainable AI recognized this opportunity by starting to employ the Shapley value for the construction of additive feature explanations which assign an importance score to each feature. The score is then interpreted as a measure of a feature's impact on a model's predicted value for some datapoint of interest, or even a model's more general behavior across multiple datapoints such as generalization performance. Thus, features are represented by the players $\mathcal{N}$ in the emerging cooperative games and the interpretation of a feature subset's worth $\nu(S)$ depends on the modeling of ν tailored to the effect $\nu(\mathcal{N})$ to be explained. We briefly revisit the fundamentals of supervised machine learning (Abu-Mostafa et al., 2012) in the following.

Supervised machine learning. For a *feature space* $\mathcal{X}$ and *target space* $\mathcal{Y}$ one usually assumes the existence of a *ground truth* $g : \mathcal{X} \to \mathcal{Y}$ that maps each *instance* $x \in \mathcal{X}$ to a *label* $y \in \mathcal{Y}$. As g is unknown, or at least only rough properties of g are presumed, it is the task of a leaner $\mathcal{Z}$ is to pick a model h, also known as *hypothesis*, from a *hypothesis space* $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ that approximates g. Given *training data* in the form of m many datapoints $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m \subseteq (\mathcal{X} \times \mathcal{Y})^m$, $\mathcal{Z}$ learns g by selecting $h \in \mathcal{H}$ which mirrors the dependency between instances and labels in $\mathcal{D}$ within the constraints imposed by $\mathcal{H}$. If the labels are produced by g, i.e. $y_i = g(x_i)$ for all $(x_i, y_i) \in \mathcal{D}$, then intuitively a good fit $h(x_i) \approx y_i$ on the training data should promise $g \approx h$ beyond $\mathcal{D}$ at first sight. Thus, g is learned from $\mathcal{D}$ and $\mathcal{Z}$ can be understood as a mapping $\mathcal{Z} : \mathcal{D} \to \mathcal{H}$. The obtained model h is then used to make predictions $h(x)$ for so far unseen instances x whose label y is not provided. Most often the data generating process incorporates *noise* leading to labels of stochastic nature such that a deterministic mapping g is not appropriate for modeling. Instead, a probabilistic dependency between instance x and label y is assumed which can be viewed as g with added heteroscedastic noise. The datapoints are now assumed to be drawn independent and identically distributed from a joint probability distribution

P over $\mathcal{X} \times \mathcal{Y}$. Since the learner is restricted to return a deterministic model h incapable of mimicking the randomness shown by the dependency between $\mathcal{X}$ and $\mathcal{Y}$, a *loss function* $\mathcal{L} : \mathcal{Y} \to \mathcal{Y}$ is used to judge how well h approximates P. For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss $\mathcal{L}(y, h(x))$ measures the error that h makes on (x, y) by comparing the ground truth label with h's prediction. Integrating the loss over the joint distribution P, one obtains h's expected error called *risk*:

$$\mathcal{R}_P(h) := \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{L}(y, h(x))\, dP(x, y)\,. \tag{3.1}$$

Risk minimization is usually the central goal of any learner $\mathcal{Z}$. In other words, $\mathcal{Z}$ is supposed to pick the optimal hypothesis $h^* \in \mathcal{H}$ with minimal expected error $\mathcal{R}_P(h^*)$. However, as P is hidden and only palpated by the datapoints in $\mathcal{D}$, one approximates the consequently inaccessible risk of h by means of its *empirical risk*

$$\mathcal{R}_{\mathcal{D}}(h) := \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(y_i, h(x_i))\,. \tag{3.2}$$

Without elaborating on pitfalls such as under- and overfitting which potentially hint at a misspecification of $\mathcal{H}$, the hypothesis with minimal empirical risk is commonly taken as a proxy for h^*. Further of importance is the shape of $\mathcal{X}$ and $\mathcal{Y}$.

- An instance $x \in \mathcal{X}$ is typically, or can always be generalized to a vector of feature values $x = (x_1, \ldots, x_n)$ such that $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ is composed of n many features. For example, the feature space of patient data could be multi-dimensional containing features such as age, blood pressure, and a binary indicator whether a patient is a smoker or not.

- The target space $\mathcal{Y}$ can either be real-valued or discrete. For sake of simplicity, we omit more sophisticated variants. The learning task for the former case is called *regression* and *classification* for the latter. Continuing the example of patient data, the target variable could be the required duration of medical attention to cure a patient, constituting a regression task, or for classification, one might want to predict the type of disease a patient suffers from.

Coming back to our alluded use case of the Shapley value, the employed models have in recent years rapidly increased in complexity, e.g. deep neural networks or boosted trees, as they demonstrate superior predictive performance and the ability to generalize intricate patterns in data. As a consequence, the relationship between instance x and predicted value $h(x)$ or other properties of h have grown

intractable to the human user or developer. Ergo, the field of explainable AI aims at turning h interpretable such that its decision-making becomes comprehensible, although making simplifications to this end. Additive feature explanations assigning feature importance scores (based on the Shapley value) are one of many methods to approach interpretability, see (Adadi and Berrada, 2018; Molnar, 2022) for a broader overview, which exhibit two appealing properties.

- The explanation is *post hoc* as it is applied on the model h after training. Hence, the training procedure performed by the learner $\mathcal{Z}$, including the choice of the hypothesis space $\mathcal{H}$ is left untouched. This allows to explain provided models without the knowledge of $\mathcal{Z}$. On the contrary, intrinsic interpretability studies how to learn and explain models that are already intrinsic by design due to their simpler structure (Molnar, 2022) such as linear models or decision trees.

- Moreover, additive feature explanations are *model-agnostic* since they do not require access to the inner workings of h and are thus capable of handling any model. This is achieved by viewing h as a black box whose behavior to input is only observed by its output, whereas other methods are tailored to specific model types and leverage the knowledge of their structure, e.g. decision trees (Lundberg et al., 2020) and neural networks (Shrikumar et al., 2017).

Further, one distinguishes between *local* and *global* explanations (Molnar, 2022).

- Local explanations consider the model's prediction $h(x)$ for a single instance x of interest. Quantifying how each feature value x_i affected $h(x)$, sheds light on how h generated the prediction. The derived *feature attribution* scores do not only hint at which features have been influential, but also which class label the contained information in x_i favors in the case of classification, or whether it had a positive or negative impact in the case of regression. However, the scores do not make a statement how even a slightly different feature value $x_i' \neq x_i$ would have changed $h(x)$ or how important a feature $\mathcal{X}_i$ is in general.

- Global explanations on the other hand, try to capture each feature's, not the feature values', effect on the model's behavior from a more holistic view, usually including multiple datapoints. A popular effect to decompose is the reduction in empirical risk $R_{\mathcal{D}'}(h)$ achieved by the features on some dataset $\mathcal{D}'$ that is not necessarily equal to $\mathcal{D}$ used for training, often disjunct, but usually generated by the same distribution P. One interprets then the resulting *feature importance* score for $\mathcal{X}_i$ as a measure of its utility for h to generalize the data.

In the following, we introduce the emerging cooperative games of Shapley-based feature explanations for both types of explanations. While the player set $\mathcal{N}$ represents all features, or their values of an instance x, throughout all examples, the value function ν is the crucial component to construct cooperative games that appropriately model the effect to be explained. Hence, ν is subject to change and can take many shapes depending on the desired explanation type and interpretation.

Local Shapley-explanations. Starting with the simpler case of regression, for given h and x it is obvious to demand $\nu(\mathcal{N}) = h(x)$ such that each feature value x_i is assigned an attribution score ϕ_i which quantifies its fair share of $\nu(\mathcal{N}) - \nu(\emptyset)$. This alone already raises two questions: How do we meaningfully define $\nu(S)$ for feature subsets $S \subset \mathcal{N}$ and how do we interpret $\nu(\emptyset)$? The construction of $\nu(S)$ should capture the prediction of h for the instance x only using the feature values S. Hence, this requires to remove the features $\mathcal{N} \setminus S$, represented by absent players, from x. In this context, the first question can be rephrased as: what does it mean for a feature to be present, or absent, in an instance x?

Unfortunately, one can not simply cut features from x as h is trained to take in an n-dimensional vector and thus unable to process an input of different dimensionality. Instead, we seek to mask the absent feature values such that their contained information is hidden from the model, simulating the removal of features. The masking, also called *feature imputation*, is performed by a *removal function*

$$f : \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \times \mathcal{P}(\mathcal{N}) \to \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \tag{3.3}$$

which manipulates an instance x to $f_S(x) := f(x, S)$ such that $h(f_S(x))$ is well-defined and can be used within the construction of ν. Since S expresses the features to remain present, their feature values should be left untouched by f, i.e. $f_S(x)_i = x_i$ for all $i \in S$. Equipped with f as a tool for masking, we can define

$$\nu(S) = h(f_S(x)) \text{ for all } S \subseteq \mathcal{N} \tag{3.4}$$

which coincides with our requirement $\nu(\mathcal{N}) = h(x)$. The worth $\nu(\emptyset)$ becomes now h's prediction with no information of x available. Still, in order to make sense of $\nu(\mathcal{N}) - \nu(\emptyset)$ and the resulting scores ϕ_i, f needs a meaningful definition. There exist multiple possibilities, each leading to a different interpretation of ϕ_i. We briefly present selected variants and refer to (Sundararajan and Najmi, 2020; Covert et al., 2021) for more extensive overviews:

Baseline imputation: Strumbelj and Kononenko (2010) substitute absent feature values by those of a baseline instance $z \in \mathcal{X}$, i.e. $f_S(x)_i = z_i$ for all $i \in \mathcal{N} \setminus S$. As a result, one obtains $\nu(\emptyset) = h(z)$ and the Shapley values decompose the prediction difference $h(x) - h(z)$ caused by the change in information from z to x. The choice of z is not trivial and could potentially complicate the interpretation as it poses the risk for mixtures of x and z to form that are out of the distribution P. Thus, h would be evaluated on instances of a shape that it has not seen before, challenging the meaningfulness of ν severely.

Marginal imputation: To circumvent the burden of having to specify a baseline instance, one can instead impute feature values by randomly drawing from each feature's marginal distribution over $\mathcal{X}_i$ conditioned on the present values from S which is derived from some dataset $\mathcal{D}'$ (Strumbelj and Kononenko, 2014; Lundberg and Lee, 2017). Now that f is not deterministic anymore, one usually takes the expectation of the resulting predictions leading to $\nu(S) = \mathbb{E}[h(f_S(x)_i)]$. The aforementioned deficiency for baseline imputation remains since each feature value is substituted independently.

Conditional imputation: Aas et al. (2021) counteract on this drawback by imputing all feature values collectively at hand of a distribution of dimension $|\mathcal{N} \setminus S|$ conditioned on the present features values. While this approach promises to construct realistic instances, Sundararajan and Najmi (2020) argue how its practicability is limited by the difficulty of eliciting the required conditional distribution over multiple features from the available data $\mathcal{D}$.

For classification tasks the real-valued effect to divide is not immediately visible. In the example of patient data, the class labels in $\mathcal{Y}$ could be names of diseases. Nevertheless, one can still construct a value function by demanding from h to output a predicted probability distribution over $\mathcal{Y}$, for which each $h_c(x)$ is interpreted as the model's estimated probability that $c \in \mathcal{Y}$ is the ground truth label of x. The multidimensionality of $h(x)$ causes a new problem to be solved as it is not immediately clear which class probability to select or how to combine them. Strumbelj and Kononenko (2010) propose class-wise value functions

$$\nu_c(S) = h_c(f_S(x)) \text{ for all } S \subseteq \mathcal{N} \tag{3.5}$$

such that one obtains an additive explanation for each class. Here it is common to only consider the cooperative game $(\mathcal{N}, \nu_{c^*})$ for the predicted label $c^* =$

$\arg\max_{c \in \mathcal{Y}} h_c(x)$. Different removal functions f can be applied orthogonally to the classifier h, as presented above for regression.

Global Shapley-explanations. For global explanations, not the prediction $h(x)$ for a specific instance but rather the model's general behavior is of interest, for example its empirical risk $\mathcal{R}_{\mathcal{D}'}$ on a separate test dataset $\mathcal{D}'$ with $\mathcal{D} \cap \mathcal{D}' = \emptyset$. As one desires to quantify each feature's contribution to h's generalization performance, one would not define $\nu(\mathcal{N}) = \mathcal{R}_{\mathcal{D}'}(h)$ but rather decompose the reduction in risk $\nu(\mathcal{N}) = \mathcal{R}_{\mathcal{D}'}(h_\emptyset) - \mathcal{R}_{\mathcal{D}'}(h)$ that the presence of all features yields. The reference risk is computed for the model $h_\emptyset$ that does not use any features. This leads us back to the question of how to perform feature removal, necessary for the construction of the complete value function ν over all feature subsets.

Cohen et al. (2007) and Pfannschmidt et al. (2016) let $\mathcal{Z}$ retrain h for each $S \subset \mathcal{N}$, already excluding the absent features $\mathcal{N} \setminus S$ from the training data $\mathcal{D}$ such that the resulting model h_S expects $|S|$-dimensional instances. Thus, they define the value function as

$$\nu(S) = \mathcal{R}_{\mathcal{D}'}(h_\emptyset) - \mathcal{R}_{\mathcal{D}'}(h_S) \text{ for all } S \subseteq \mathcal{N}\,. \tag{3.6}$$

The worth of a feature subset can be interpreted as the gained predictive performance caused by the inclusion of these features. One might wonder, what shape $h_\emptyset$ takes. For regression with the mean squared error it predicts the mean target value observed in $\mathcal{D}$ and the majority class for classification with 0-1 loss. Notably, this construction is rather a statement about the learner $\mathcal{Z}$ than the initial model since h is only involved in $\nu(\mathcal{N})$, assuming that in fact $h_\mathcal{N} = h$. Nevertheless, it constitutes a meaningful way of measuring each feature's importance.

In order to stick to the provided model h and not require the involvement of $\mathcal{Z}$, Covert et al. (2020) propose to apply feature masking as done for local explanations. In particular, a removal function f based on conditional imputation is used to construct

$$\nu(S) = \mathbb{E}\left[\mathcal{R}_{\mathcal{D}'}(f_\emptyset(h))\right] - \mathbb{E}\left[\mathcal{R}_{\mathcal{D}'}(f_S(h))\right]\,, \tag{3.7}$$

where the randomness is taken w.r.t. the imputation. Inspired by (Heskes et al., 2020) for local explanations, Breuer et al. (2024) go one step further and involve causal relationships between features within the imputation performed by f.

3.2 Feature Explanations with Shapley Interactions

Additive feature explanations exhibit the appeal of high readability due to their low complexity, being specified by only one value per feature. Meanwhile, other works have investigated their validity (Kumar et al., 2020; Kumar et al., 2021; Gosiewska and Biecek, 2020) and observed that the additive decomposition is mostly too simple to grasp non-additive structures that complex models possess, as they would otherwise not be concerned for explanation. From the perspective of a cooperative game, it is self-understood that n many importance scores can hardly represent a value function assigning worth to 2^n many feature subsets without making drastic simplifications. Kumar et al. (2020) argue how the additivity axiom conceptually limits the Shapley value, rendering it inappropriate to faithfully capture the intricate dependencies between features that arise so often in practice. Further, Gosiewska and Biecek (2020) showcase that the obliviousness to these dependencies potentially leads to inconsistent explanations that users can not rely on.

The aforementioned works collectively hint at interactions between features to overcome the shortcomings of the Shapley value. Fittingly, cooperative game theory offers with the notion of Shapley interactions (see Section 2.3) already a tool to unveil the latent synergies of higher dimensionality between features. The Shapley interactions can be derived without the need to alter the constructed cooperative games for Shapley-based explanations. A gained benefit of the shared origin in form of the same cooperative game is that the Shapley values do not have to be dismissed but rather extended by the obtained interactions. In other words, Shapley interactions do not replace additive explanations but enrich them with additional information. As the dimensionality of the interactions to be considered can be specified by the human user himself, Shapley interactions give room to individually adjust how the inherent trade-off between fidelity and readability is tackled by tuning the complexity of the explanation.

Whereas, at the example of local explanations for regression, a negative Shapley value ϕ_i indicates that the feature value x_i contributes to the reduction of the predicted value $h(x)$, the interpretation of Shapley interactions judging by the sign alone is not quite as clear. In the context of feature explanations, one can distinguish between three ways in which the pairwise interaction $I^{\phi}_{i,j}$ of two features x_i and x_j having positive Shapley values ϕ_i and ϕ_j impacts the prediction. A positive sign is commonly interpreted as evidence that x_i and x_j complement each other such that the combination of both provides additional information, increasing $h(x)$ even

further than the features do on their own. A negative sign might hint at how the combination of x_i and x_j reveals new information to the model that speaks in favor of a reduction of $h(x)$, opposite to the feature values' individual impact. However, this is not necessarily the case. As a third possibility, taking into account both x_i and x_j could still increase $h(x)$ but in sub-additive manner compared to the individual contributions of x_i and x_j, leading to $I^{\phi}_{i,j}$ being negative. For this phenomena, the features are often said to contain redundant information. The incentive for h to further increase its prediction is reduced by the information already provided by one feature, that the other at least partly shares.

3.3 Selection of Machine Learning Entities

So far, we have considered feature importance scores on the basis of the Shapley value for the purpose of explaining machine learning models. However, these scores are not limited to being only applied within explainability, but are also suitable to serve as a tool to more tangible goal-driven tasks such as feature subset selection. Cohen et al. (2007), Pfannschmidt et al. (2016), Becker and Bengs (2023), and Sebastián and González-Guillén (2024) evaluate the usefulness of each feature by means of Shapley values as they would be derived for global explanations, and select features with the highest assigned scores. This application entails a further incentive to consider the top-k identification (see Section 2.4 and Section 4.5) if one wants to preselect a feature subset of fixed size k based on Shapley values.

Selecting a subset of the most useful features before the learning procedure promises to achieve better generalization performance when complex models show the inclination to overfit on the given training data. Moreover, having to process fewer feature values reduces computational expenditures caused by training and inference. Last but not least, a reduction of the feature space itself can already aid interpretability as the influences of fewer features have to be overseen. Beyond supervised learning, Balestra et al. (2022) recognize the possibility of constructing cooperative games out of unlabeled data, only containing instances, such that the resulting Shapley values reflect the features' importance within that point cloud. Feature selection can then be likewise conducted to mitigate the curse of dimensionality.

Actually, the versatility of cooperative games opens the door for an even wider extension of the Shapley value's range of applications. The games in Section 3.1 are constructed by viewing the value function ν as the tool for appropriate modeling

while keeping the player set $\mathcal{N}$ fixed to represent features. Interestingly, an observed effect such as a model's empirical risk does not need to be distributed among features. Instead, various other entities involved in the learning process, for example the datapoints used for training or structural components of the model, can constitute the players of a cooperative game. We refer to (Rozemberczki et al., 2022) for a broader overview.

Data valuation and federated learning. The field of data valuation quantifies the value of each datapoint in the training set for learning and cooperative games are applied by formulating each datapoint $(x_i, y_i) \in \mathcal{D}$ as a player. In order to let the Shapely value reflect how much a datapoint contributes to the generalization performance, the value function is defined to map each coalition S to a performance score, for example test accuracy for classification, of the model h_S that the learner returns when using only $S \subseteq \mathcal{D}$ as the training set (Ghorbani and James Y. Zou, 2019; Jia et al., 2019; Wu et al., 2023). Sorting the datapoints by their Shapley values and successively removing those of least importance facilitates a flexible approach to reduce the training set. Further, excluding datapoints with negative Shapley values constitutes an approach to remove harmful outliers in the data, which potentially mitigate predictive performance. Another economically motivated purpose is to pay out data owners proportionally to the Shapley value of each datapoint they provided to construct a richer collective training set. More generally, the branch of federated learning studies how multiple clients with their own private data can collaborate by offering their datasets to train a model in decentralized manner. Viewing each client as a player, the Shapley value poses a mechanism to fairly payout clients according to the gain in predictive performance of the joint model that their data provides (Liu et al., 2022; Sim et al., 2020).

Model pruning. The complexity of modern machine learning models is not only rooted in highly non-linear operations that are performed during inference but also in the increasing number of structural components that these models are made of. Common examples are deep neural networks such as *BERT* (Devlin et al., 2019) for natural language processing with millions of parameters and thousands of artificial neurons, or random forests aggregating dozens if not hundreds of decision trees. Hence, the possibility of assigning importance to individual components has attracted interest for multiple reasons. First, the sheer ability to assess the impact of each component on the generalization performance may provide an understanding that could potentially guide model design or more precisely the

design of the hypothesis space $\mathcal{H}$. Similarly to feature selection, reducing the number of components, known as *model pruning*, promises to mitigate the threat of overfitting caused by too expressive models coupled with sparse training data (LeCun et al., 1989). The reduction in size at the expense of the model's expressivity caused by pruning is accepted as a compromise to render large models practicable to a wider range of users and application scenarios with constrained computational resources. Ghorbani and James Y Zou (2020) formulate neurons as players to prune deep neural networks by removing neurons whose Shapley values indicate a low or even harmful contribution to a performance score of the network. Rozemberczki and Sarkar (2021) prune model ensembles that aggregate multiple base models to a bigger model. In particular, random forests consisting of decision trees are considered by defining the player set as the set of initially contained decision trees.

Including interactions. Sorting entities such as features, datapoints, or neurons by their Shapley values and discarding the assumably least important ones is a conceptually straightforward approach. However, it is not free of pitfalls. Entities that share contributing factors, for example features with similar information to improve predictive performance, possess a certain redundancy which is not reflected by the Shapley value. In fact, for two correlated features $\mathcal{X}_i$ and $\mathcal{X}_j$ having relatively high Shapley values, including $\mathcal{X}_i$ might render $\mathcal{X}_j$ barely useful, as $\mathcal{X}_i$ already provides most of the information that $\mathcal{X}_j$ could contribute. In the context of pruning, one might also be fooled by the sorting in the opposite way. Multiple mutually correlated features can share an importance mass via their Shapley values that one feature alone would inherit if the others were absent. Thus, potentially all of these features seem in comparison to those not sharing importance mass relatively insignificant and are at the danger of being pruned, although at least one of them would make a significant contribution. Whereas the importance of an entity is falsely boosted in the first example, it is hidden by the Shapley value in the second example. To avoid falling victim to these pitfalls, Chu and Chan (2020) propose in the context of feature subset selection the incorporation of interaction between features to adjust the sorting. As discussed in Section 3.2, Shapley interactions bear the potential to expose the redundancy between features since their core building block, the discrete derivative, measures the degree to which one player's presence impacts the other player's contribution. Obviously, this idea can be generalized to any kind of entity represented by a player within a cooperative game.

Since the versatility of cooperative games enables the Shapley value and interactions to be applied for a multitude of diverse tasks, we see ourselves confirmed in the

pursuit of approximation algorithms that are universally applicable, independent of the actual domain being modeled by the game. Further, we view the employment of tailored heuristics within the value function, capturing how particular entities such as features or datapoints tend to behave and shape a coalition's worth, as subordinate to the research on approximation techniques yielding a general understanding and possibly theoretical guarantees that are not restricted by specific assumptions. We meet this motivation of domain-independent methods in the following chapter.

4 Contribution and State of the Art

The contribution of this thesis is divided into four parts. We give a brief overview by concisely presenting each part in the following. After a categorization of selected approximation methods for the Shapley value in Section 4.1, providing the necessary background of state-of-the-art literature, Section 4.2 to 4.5 elaborate on each contribution part in more detail.

Contribution (I)-(III): Shapley Value Approximation via Stratification. To begin with, we detach in (I) the Shapley value from the popular notion of marginal contributions on which it is built on, and propose a different representation relying on the aggregation of single coalition values as the foundation of our developed approximation method. Combining it with stratification as a variance reduction technique for mean estimates facilitates the integration of a more efficient update rule that utilizes each sampled coalition to update the estimate of each player. In addition, the derived theoretical guarantees in (I) result in a twofold benefit: they allow the construction of confidence intervals around the estimates and give insight on which games our algorithm is advantageous. Further, we refine our method in (II) by observing that the allocation of samples across the strata plays a vital role in the reduction of the approximation error and exhibits room for optimization. We transfer adaptive sampling from stratifying methods based on marginal contributions to our approach such that the employed allocation is no longer given a priori but updated during the approximation process itself. In addition to the empirical improvement, we investigate in (II) stratification also from a more conceptual viewpoint, quantify its theoretical potential at the hand of the optimal allocation in hindsight, and answer to which degree our method closes this gap. At the example of feature explanations for unlabeled data based on total correlation, we illustrate in (III) the efficacy of stratifying methods, concluding that these leverage the relation between a coalition's size and its worth to outshine competing methods.

Contribution (IV): Shapley Value Approximation via Optimization. We tackle the approximation problem with an alternative method in (IV) by deviating from the common perspective of the Shapley value as an expected value of marginal contributions or coalition values. Instead, we construct a structured and parameterizable surrogate game that mimics the cooperative game at hand by fitting its value function to that of the given game after observing sampled coalitions. The key idea is to exploit the surrogate game's structure and compute its exact Shapley values in polynomial time. At the same time, a sufficient degree of flexibility in that structure promises a good fit such that the surrogate game's Shapley values serve as precise estimates for that of the given game. Choosing the value function to be k-additive allows us to extract its Shapley values immediately as they are directly contained within this representation. Further, we provide as an analytical result that the fitting of a k-additive game forces its Shapley values to exactly match those of the given cooperative game after having observed the latter in its entirety. We interpret this result as proof to the theoretical soundness of our approach, implying that the error of the retrieved estimates converges to zero during the approximation process.

Contribution (V): Approximation of Shapley Interactions. In the spirit of (I), we present in (V) a representation of Shapley interactions that emphasizes coalition values instead of discrete derivatives and also stratify it by the coalitions' size. So to speak, this work serves as an extension and methodological transfer of the approach presented in (I) from the special case of the Shapley value to the superordinate class of cardinal interaction indices. Moreover, this work exhibits a unifying nature: all semi-values and cardinal interaction indices can be approximated simultaneously during the sampling process or afterwards without the requirement to specify indices of interest in advance. This is made feasible by the observation that the maintained strata are atomic building blocks shared by all semi-values and cardinal interaction indices, which only differ in the weighting of these strata.

Contribution (VI)-(VII): Top-k Shapley Players Identification. We recognize in (VI) how the task of identifying the k players with highest Shapley values differs substantially enough from that of approximating all values precisely to employ different algorithmic approaches. Most importantly, since the numerical estimates do not need to be precise across the whole player set, the identification problem offers the opportunity to sacrifice precision of estimates whose players' top-k membership can be determined with confidence in favor of those who are metaphorically speaking located close to the top-k border. This allows to redistribute available samples from

the former to the latter type of players. Based on this observation, we establish a novel connection to multi-armed bandits from the field of online machine learning, viewing each player as an arm of a slot machine that can be pulled to obtain a sample from its distribution of marginal contributions as a reward. Particularly for the top-k identification problem, we discover in (VII) the utility of comparing marginal contributions to the same reference coalition between players, leading to the variance reduction technique of antithetic sampling based on correlated observations. Motivated by this insight, we develop an approximation algorithm that exploits a new representation of the Shapley value utilizing an altered notion of marginal contribution and incorporates techniques inspired by multi-armed bandits. Moreover, we highlight the difference to the approximate-all problem by comparing the performance of algorithms on both tasks and observing that approximation quality is not necessarily transferable between them.

4.1 Categorization of Approximation Methods

The Shapley value's approximation problem has fostered the usage of varying techniques for estimation fueled by its richness in representations. Subsequently, a diverse landscape of proposed methods has formed within the literature. As we focus on methods that are applicable to any cooperative game without posing structural assumptions that rely on the context of a specific domain, we give in the following a brief overview on the taxonomy of such approximation algorithms. Figure 4.1 illustrates this taxonomy and incorporates contributions of this thesis. Within the plethora of proposed methods, we distinguish between two main classes.

Approximation through mean estimation. The first class interprets the Shapley value as an expected value and thus encompasses approximation algorithms that conduct mean estimation. Most popular among those is to view the Shapley value of each player as its expected marginal contribution w.r.t. a discrete distribution in which each marginal contribution is assigned its weight in Definition 2.14 as its probability. *ApproShapley* (Castro et al., 2009) realizes sampling from this distribution by drawing random permutations of the player set and using the appearing marginal contributions in these sequences as samples. It relies on the Shapley value's representation based on permutations given by Equation 2.3 and draws the same number m of samples $\Delta_i\left(S_i^{(1)}\right), \ldots, \Delta_i\left(S_i^{(m)}\right)$ for each player since every player

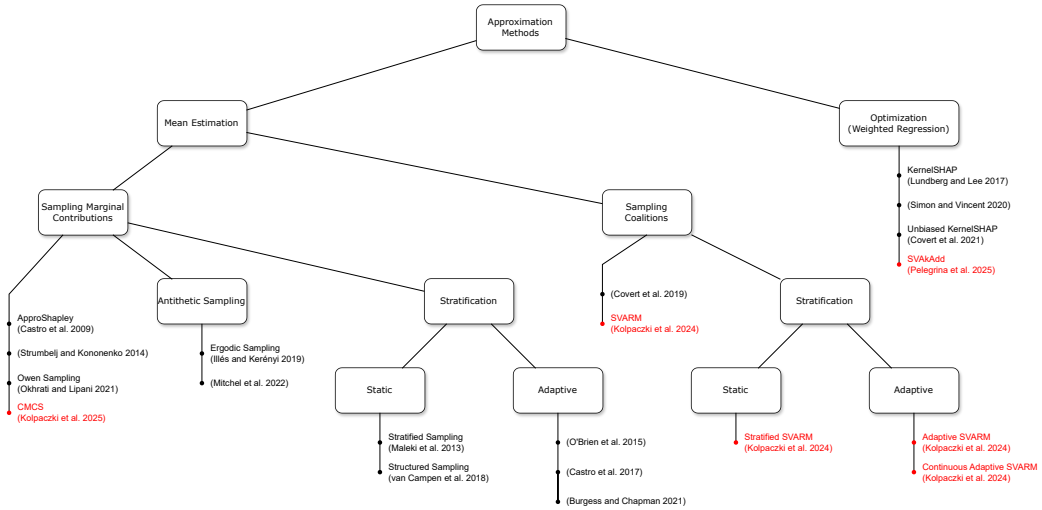


Figure 4.1.: Taxonomy of selected domain-agnostic approximation algorithms for the Shapley value. Selected contributions of this thesis are marked in red.

exhibits exactly one marginal contribution in a permutation. Hence, each estimate is computed as

$$\hat{\phi}_i = \frac{1}{m} \sum_{m'=1}^{m} \Delta_i \left(S_i^{(m')} \right) . \tag{4.1}$$

This results in an MSE (see Equation 2.11) of

$$\frac{1}{nm} \sum_{i=1}^{n} \sigma_i^2 , \tag{4.2}$$

where $\sigma_i^2 := \mathbb{V}[\Delta_i(S)]$ is the variance of player i's marginal contributions for random $S \subseteq \mathcal{N} \setminus \{i\}$ with distribution $\mathbb{P}(S) = \frac{1}{n \cdot \binom{n-1}{|S|}}$. Drawing, the marginal contributions independently between the players would consume two budget tokens per draw, leading to a maximum of $m \leq \frac{T}{2n}$ many samples per player. Instead, by reusing the already evaluated coalitions in a permutation, this upper bound can be increased to $m \leq \frac{n}{n+1} T$ since only $n + 1$ evaluations have to made per permutation. Despite the fixed budget T, this improvement in terms of higher m leads to a reduction in MSE.

In contrast, Strumbelj and Kononenko (2014) propose an improvement by sequentially choosing to sample the next marginal contribution from that player whose contribution to the MSE would reduce the most, taking into account the sample variance $\hat{\sigma}_i^2$ of its observed marginal contributions and the hitherto number of samples m_i'. In particular, the player i maximizing

$$\frac{\hat{\sigma}_i^2}{m_i'} - \frac{\hat{\sigma}_i^2}{m_i' + 1} \tag{4.3}$$

is greedily selected. The new key element here is that the approximation algorithm is *adaptive* in the sense that it adjusts its sampling procedure to the observations collected so far, whereas *ApproShapley* is *static*. As a consequence each player is assigned its own total number of samples m_i instead of a uniform m.

Maleki et al. (2013) introduce the variance reduction technique of stratification to the field of Shapley value estimation. Each player's population of marginal contributions $\Delta_i(S)$ is partitioned by grouping them into n many strata sharing the same size $|S|$. For our convenience, this partitioning can also be expressed algebraically, describing the Shapley value as an average of strata values $\phi_{i,\ell}$ which in turn are arithmetic means of their contained marginal contributions:

$$\phi_i = \frac{1}{n} \sum_{\ell=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \{i\} \\ |S| = \ell}} \Delta_i(S)}_{=:\phi_{i,\ell}} . \tag{4.4}$$

The samples are taken independently from each stratum such that the resulting subestimates $\hat{\phi}_{i,\ell}$ are aggregated back to

$$\hat{\phi}_i = \frac{1}{n} \sum_{\ell=0}^{n-1} \hat{\phi}_{i,\ell} . \tag{4.5}$$

An allocation that assigns $m_{i,\ell} \geq 1$ samples to each stratum results in an MSE of

$$\mathbb{E}[\text{MSE}] = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{\ell=1}^{n-2} \frac{\sigma_{i,\ell}^2}{m_{i,\ell}} \tag{4.6}$$

with stratum variance $\sigma_{i,\ell}^2 := \mathbb{V}[\Delta_i(S)]$ for uniformly distributed $S \subseteq \mathcal{N} \setminus \{i\}$ of size ℓ. Note that the stratum variances are constants of the cooperative game but the allocation $m_{i,\ell}$ is free of choice. Obviously, it holds $\sigma_{i,0}^2 = \sigma_{i,n-1}^2 = 0$ for all players and retrieving the single marginal contributions of the respective strata suffices to exclude these from further consideration, requiring only a budget of $2n + 2$. Hence, in search of minimal MSE, Equation 4.6 forms in combination with the budget constraint $\bar{T} := T - 2n - 2$ an optimization problem. Its solution, the optimal choice of $m_{i,\ell}$, is known as the Neyman allocation (Neyman, 1934), taking the shape of

$$m_{i,\ell}^* = \frac{\sigma_{i,\ell}}{2 \sum_{j=1}^{n} \sum_{k=1}^{n-2} \sigma_{j,k}} \cdot \bar{T} , \tag{4.7}$$

where we for sake of simplicity omit the constraint that each $m_{i,\ell}$ must be a natural number. Strata with higher standard deviation $\sigma_{i,\ell}$ require proportionally more samples in the Neyman allocation. Since the stratum variances (or standard deviations) are a priori unknown, a *static* allocation as employed by *Stratified Sampling* (Maleki et al., 2013) has no chance of reaching it and thus minimizing the MSE.

To combat this lack of knowledge, adaptive methods adjust their sampling accordingly by favoring strata that exhibit higher observed variance, thus impact the MSE to greater degree. Castro et al. (2017) propose a two-phased algorithm that first explores the cooperative game by equifrequently sampling from all strata. Based on the estimated variances $\hat{\sigma}^2_{i,\ell}$, an estimate of $m^*_{i,\ell}$ is used to chase the Neyman allocation in the subsequent exploitation phase. O'Brien et al. (2015) take inspiration from multi-armed bandits to smoothly transition from exploration (collecting evenly samples from all strata) to exploitation (mimicking the optimal allocation) with the help of a sigmoid function, instead of abruptly switching phases. Burgess and Chapman (2021) employ the *Stratified Empirical Bernstein Bound* to identify the next player to sample for that promises maximal MSE reduction. Despite their appeal to optimally employ stratification, none of these adaptive methods come with guarantees for the approximation quality in the fixed-budget setting. So far, only their unbiasedness has been shown.

Owen Sampling (Okhrati and Lipani, 2020) can remotely be categorized as a stratifying method but is built upon a quasi-continuous distribution of marginal contributions. Instead of viewing the Shapley value as a discrete sum, Owen (1972) shows that it can be represented by the integral

$$\phi_i = \int_0^1 \mathbb{E}\left[\Delta_i(S_{i,q})\right] \, dq \tag{4.8}$$

where $S_{i,q} \subseteq \mathcal{N} \setminus \{i\}$ is randomly constructed: a biased coin toss decides for each player $j \in \mathcal{N} \setminus \{i\}$ whether it is present in $S_{i,q}$ or not. So the probability that $S_{i,q}$ equals some $S \subseteq \mathcal{N} \setminus \{i\}$ is

$$\mathbb{P}(S_{i,q} = S) = q^{|S|} \cdot (1-q)^{n-|S|-1} \,. \tag{4.9}$$

Okhrati and Lipani (2020) numerically integrate Equation 4.8 by palpating it at fixed equidistant points $q \in [0,1]$ and estimating the expected marginal contribution for each q through sampling. This representation reveals an intricate connection leading back to Definition 2.14. Drawing a random $q \in [0,1]$ and next a marginal contribu-

tion according to Equation 4.9 is equal to drawing each marginal contribution with its weight within the Shapley value.

All previous methods have in common that marginal contributions are drawn independently for each player, but not necessarily between them, e.g. *ApproShapley*. By drawing dependent observations, Illés and Kerényi (2019) and Mitchell et al. (2022) apply the technique of antithetic sampling which exploits covariances between samples to decrease each estimate's variance. In light of the bias-variance decomposition (see Equation 2.12), this promises to further reduce the MSE.

Another subbranch of approximation methods conducting mean estimation foregoes the notion of marginal contributions and splits the Shapley value into two sums:

$$\phi_i = \underbrace{\sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot \nu(S \cup \{i\})}_{=:\phi_i^+} - \underbrace{\sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot \nu(S)}_{=:\phi_i^-} . \tag{4.10}$$

Covert et al. (2019) implicitly adopt this view and instead of sampling for each sum separately, each drawn coalition is used to update an estimate of ϕ_i^+ and ϕ_i^- via importance sampling. Wang and Jia (2023) estimate the Banzhaf value in equivalent fashion without importance sampling due to its uniform weights.

Approximation through optimization. Although mean estimation might be considered as the most natural way to approximate the Shapley value given its shape of a weighted sum, the branch of optimization constitutes a popular alternative. These methods follow a completely different approach by not estimating the Shapley values ϕ_i of a given game $(\mathcal{N}, \nu)$ directly, but instead compute the exact Shapley values ϕ_i' of a surrogate game $(\mathcal{N}, \nu')$ that is to be close to $(\mathcal{N}, \nu)$. The intuition behind this idea is that a value function ν' similar to ν should also yield similar Shapley values and thus precise estimates. This approach is fruitful under two conditions: (i) the surrogate game's Shapley values ϕ' quickly converge to ϕ and (ii) its Shapley values can be computed in polynomial time. The second condition is met by imposing a class of highly structured value functions for whom closed-form solutions exist. At the same time, incorporating a certain flexibility into ν' is desirable such that it can reflect the shape of ν. Given observed coalition values $\nu(S_1), \ldots, \nu(S_T)$ through sampling without replacement, the surrogate game is fitted to minimize an

objective function that quantifies the dissimilarity between ν and ν'. In particular, the objective function most often emulates weighted regression in the form of

$$\sum_{t=1}^{T} w_{S_t} \left(\nu'(S_t) - \nu(S_t)\right)^2 . \tag{4.11}$$

On this basis, *KernelSHAP* (Lundberg and Lee, 2017) employs a simplistic yet effective surrogate game in which each player $i \in \mathcal{N}$ possesses a coefficient c_i that is added to the worth of a coalition upon i's inclusion, leading to the value function

$$\nu'(S) = c_0 + \sum_{i \in S} c_i \quad \text{for all } S \subseteq \mathcal{N} , \tag{4.12}$$

where c_0 is used to handle games in which $\nu(\emptyset) \neq 0$. Hence, within the surrogate game all players are dummy players. Using weights $w_S = \binom{n-2}{|S|-1}^{-1}$ for all $S \in \mathcal{P}(\mathcal{N}) \setminus \{\emptyset, \mathcal{N}\}$ and adding the efficiency constraint (see Definition 2.12), estimates ϕ'_i are obtained by solving the following optimization problem:

$$\begin{aligned} \min_{c_0,\ldots,c_n} \; & \sum_{t=1}^{T} \frac{1}{\binom{n-2}{|S|-1}} \left(\nu'(S_t) - \nu(S_t)\right)^2 \\ \text{s.t.} \; & \sum_{i=1}^{n} c_i = \nu(\mathcal{N}) - \nu(\emptyset) \end{aligned} \tag{4.13}$$

Despite the simplicity of ν', the solution to Equation 4.13 yields the Shapley values of $(\mathcal{N}, \nu)$ when all coalitions except for $\emptyset$ and $\mathcal{N}$ are included in the objective function (Charnes et al., 1988). This implies the convergence of ϕ'_i to ϕ_i during approximation for an increasing number of samples and is thus a vital step towards condition (i). Fittingly, the representation of ν' given by Equation 4.12 already elicits the Shapley values of the surrogate game because $\phi'_i = c_i$ holds due to the dummy axiom, satisfying condition (ii).

In opposition to mean-estimation methods, Covert and Lee (2021) recognize the analytical difficulty to make statements about *KernelSHAP*'s properties such as bias and variance. Instead, they empirically show a non-zero bias and further propose *Unbiased KernelSHAP* that trades zero bias for higher variance. Fumagalli et al. (2023) show how this variant of *KernelSHAP* effectively coincides with mean-estimation of the Shapley value as a weighted sum of coalition values, bridging the two main branches of approximation. In another variant, Simon and Vincent (2020) tackle the convex optimization problem by performing stochastic gradient descent.

Approximation of Shapley interactions. Compared to the richness of methods for estimating Shapley values, the choice of approximation algorithms for Shapley interaction is quite sparse. Sundararajan et al. (2020) extend the Monte Carlo method of *ApproShapley* to sample discrete derivatives and estimate the Shapley interaction index as a mean discrete derivative, which is likewise applicable to the general class of cardinal interaction indices. Meanwhile, they propose the *Shapley-Taylor index* that measures interactions for order 1 up to a specified k. The Shapley-Taylor interaction for the top-order k is as well a cardinal interaction index. Tsai et al. (2023) introduce the *Faithful Shapley interaction index* measuring interactions between order 1 and k inspired by the weighted regression formed by Equation 4.11 and 4.12 but extended to interactions. Thus, their index can also be approximated via optimization. The representation of the Shapley interaction index for pairs as a solution to a weighted regression problem, similar to *KernelSHAP*, has been recently discovered under *KernelSHAP-IQ* (Fumagalli et al., 2024).

4.2 Shapley Value Approximation via Stratification

Contribution (I). We start by explicitly introducing the new representation of two sums in Equation 4.10 as a basis for approximating the Shapley value via mean estimation in absence of marginal contributions. Our result stating the non-existence of a distribution over $\mathcal{P}(\mathcal{N})$ to sample from such that ϕ_i^+ and ϕ_i^- can be estimated without bias and importance sampling, motivates us to propose *Shapley Value Approximation without Requesting Marginals (SVARM)*. It relies on the observation that each observed $\nu(S)$ can be used to update the estimate $\hat{\phi}_i^+$ for all $i \in S$ or $\hat{\phi}_i^-$ for all $i \notin S$ if drawn according to the distributions

$$P^+(S) := \frac{1}{|S|\binom{n}{|S|}H_n} \qquad \forall S \in \mathcal{P}(\mathcal{N}) \setminus \{\emptyset\}\,, \tag{4.14}$$

$$P^-(S) := \frac{1}{(n-|S|)\binom{n}{|S|}H_n} \qquad \forall S \in \mathcal{P}(\mathcal{N}) \setminus \{\mathcal{N}\}\,, \tag{4.15}$$

respectively, where H_n denotes the n-th harmonic number. This allows to not only update a single estimate for some player but all affected players depending on which distribution is being used. Hence, *SVARM* alternates between estimating ϕ_i^+-values and ϕ_i^--values and reuses each evaluated worth $\nu(S)$ for multiple players. In fact, both distributions cause with each draw $\frac{n}{H_n}$ many updates in expectation such that our combination of sampling mechanism and update rule exhibits higher budget

efficiency. In comparison, the branch of methods drawing marginal contributions can only utilize each $\Delta_i(S)$ for $\hat{\phi}_i$ as it appears in no other player's Shapley value.

In order to maximize the reusage of costly acquired samples, we incorporate the variance reduction technique of stratification and propose *Stratified SVARM* building on the representation of the Shapley value as

$$\phi_i = \frac{1}{n}\sum_{\ell=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S\subseteq\mathcal{N}\setminus\{i\} \\ |S|=\ell}} \nu(S\cup\{i\})}_{=:\phi_{i,\ell}^{+}} - \frac{1}{n}\sum_{\ell=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S\subseteq\mathcal{N}\setminus\{i\} \\ |S|=\ell}} \nu(S)}_{=:\phi_{i,\ell}^{-}} . \tag{4.16}$$

Worth mentioning is how it coincides with the proposal of (Ancona et al., 2019), although there not being used for domain-agnostic approximation as we intend. On one hand, stratifying each player's two sums promises to obtain faster converging strata estimates (see contribution (II) in Chapter 6 for a more detailed explanation). On the other, it opens up the opportunity to increase budget efficiency to a point where it reaches the *maximum sample reuse principle* (Wang and Jia, 2023), stating (in the context of the Banzhaf value) that each coalition's worth should be used for all player's estimates. Stratification allows to circumvent the difficulty of different weights between ϕ_i^+ and ϕ_i^- in the case of the Shapley value. Constructing each Shapley value estimate as a combination of estimates for the strata values, i.e.

$$\hat{\phi}_i = \frac{1}{n}\sum_{\ell=0}^{n-1} \hat{\phi}_{i,\ell}^{+} - \hat{\phi}_{i,\ell}^{-}\,, \tag{4.17}$$

makes updating all $\hat{\phi}_i$ with each sampled $\nu(S)$ feasible. The key observation here is that regardless which coalition S one considers, for each player i there exists exactly one size ℓ such that either $\phi_{i,\ell}^+$ or $\phi_{i,\ell}^-$ contains $\nu(S)$. Since these are arithmetic means, updating is straightforward without the necessity of importance sampling, thus facilitating the derivation of theoretical guarantees for the approximation quality in the fixed-budget setting. Note that in principle, *Stratified SVARM* can approximate all semivalues simultaneously by reweighting the strata value estimates in Equation 4.17. This is not even required to happen during sampling but can be performed at termination for any chosen semivalue.

The benefit gained by maximum sample reuse becomes tangible upon inspecting the MSE. If all coalitions of size $0, 1, n-1$, and n are evaluated by a warm-up phase in advance and m_ℓ coalitions of size ℓ are drawn in total, then for stratum variances

$\sigma^2_{i,\ell,+} := \mathbb{V}[\nu(S_{i,\ell} \cup \{i\})]$ and $\sigma^2_{i,\ell,-} := \mathbb{V}[\nu(S_{i,\ell})]$ with uniformly distributed random coalition $S_{i,\ell} \subseteq \mathcal{N} \setminus \{i\}$ of size ℓ, the MSE of *Stratified SVARM* is bounded by

$$\mathbb{E}[\text{MSE}] \leq \frac{1}{n^2} \sum_{\ell=2}^{n-2} \frac{1}{m_\ell} \sum_{i=1}^{n} \frac{\sigma^2_{i,\ell-1,+}}{\ell} + \frac{\sigma^2_{i,\ell,-}}{n-\ell}\,. \tag{4.18}$$

We specify m_ℓ appropriately, in other words dictating how many coalitions of size ℓ are to be drawn, yet not including any knowledge of the stratum variances, to simplify the bound and derive

$$\mathbb{E}[\text{MSE}] \leq \frac{2 \log n}{n^2 \bar{T}} \sum_{i=1}^{n} \sum_{\ell=2}^{n-2} \sigma^2_{i,\ell-1,+} + \sigma^2_{i,\ell,-} \in \mathcal{O}\left(\frac{\sigma^2_{\max_{+/-}} \cdot \log n}{\bar{T}} \right), \tag{4.19}$$

where $\bar{T} = T - 2n - 2$ is the remaining budget after evaluating all coalitions of size $0, 1, n-1$, and n. Here, $\sigma^2_{\max_{+/-}} := \max_{i \in \mathcal{N}, \ell \in \{2,\dots,n-2\}} \{\sigma^2_{i,\ell-1,+}, \sigma^2_{i,\ell,-}\}$ is a constant of the game itself. We would like to stress the point that except for the warm-up phase *Stratified SVARM* achieves an MSE that grows sublinearly with the number of players if one considers the maximum stratum variance to be independent of n. For completeness, sampling from all sizes in equal frequencies yields the same asymptotic behavior. In comparison, the asymptotic MSE resulting from stratified sampling of marginal contributions (see Equation 4.6) with equal frequencies over all sizes exhibits a worse dependency:

$$\mathbb{E}[\text{MSE}] \leq \frac{1}{n\bar{T}} \sum_{i=1}^{n} \sum_{\ell=1}^{n-2} \sigma^2_{i,\ell} \in \mathcal{O}\left(\frac{\sigma^2_{\max} \cdot n}{\bar{T}} \right), \tag{4.20}$$

where $\sigma^2_{\max} := \max_{i \in \mathcal{N}, \ell \in \{1,\dots,n-2\}} \sigma^2_{i,\ell}$. Assuming the maximum stratum variances to be equal, *Stratified SVARM* achieves an MSE reduction of a factor of roughly $\frac{n}{\log n}$ neglecting constants which corroborates its increase in budget efficiency. The experiments conducted by Muschalik et al. (2024) speak in favor of our method, stating its competitiveness across various types of cooperative games that appear in the field of machine learning.

Contribution (II). Since the sample allocation m_ℓ over strata offers a certain degree of freedom, one might pose the question of how much *Stratified SVARM* can be improved by optimizing its allocation. Here, optimization entails the adjustment to the a priori unknown stratum variances. To shed light on the potential of

stratification (for fixed partitioning into strata), we derive similar to (Neyman, 1934) the optimal sample allocation for *Stratified SVARM* to be

$$m_\ell^* = \frac{\sqrt{\sum_{i=1}^{n} \frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}}}{\sum_{k=2}^{n-2} \sqrt{\sum_{i=1}^{n} \frac{\sigma_{i,k-1,+}^2}{k} + \frac{\sigma_{i,k,-}^2}{n-k}}} \cdot \bar{T} \,. \tag{4.21}$$

While m_ℓ^* remains unknown in practice, it serves the purpose of quantifying the minimal MSE that *Stratified SVARM* could theoretically achieve. Plugging the optimal allocation into Equation 4.18 yields

$$\mathbb{E}[\text{MSE}] \leq \frac{1}{n^2 \bar{T}} \left(\sum_{\ell=2}^{n-2} \sqrt{\sum_{i=1}^{n} \frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}} \right)^2 \in \mathcal{O}\left(\frac{\sigma_{\max_{+/-}}^2}{\bar{T}} \right) . \tag{4.22}$$

Proof of asymptotics:

$$\begin{aligned}
\frac{1}{n^2 \bar{T}} \left(\sum_{\ell=2}^{n-2} \sqrt{\sum_{i=1}^{n} \frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}} \right)^2 &\leq \frac{\sigma_{\max_{+/-}}^2}{n\bar{T}} \left(\sum_{\ell=2}^{n-2} \frac{1}{\sqrt{\ell}} + \frac{1}{\sqrt{n-\ell}} \right)^2 \\
&\leq \frac{4\sigma_{\max_{+/-}}^2}{n\bar{T}} \left(\sum_{\ell=1}^{n} \frac{1}{\sqrt{\ell}} \right)^2 \\
&\leq \frac{4\sigma_{\max_{+/-}}^2}{n\bar{T}} \left(1 + \int_1^n \frac{1}{\sqrt{\ell}} \, d\ell \right)^2 \\
&\leq \frac{16\sigma_{\max_{+/-}}^2}{\bar{T}} \in \mathcal{O}\left(\frac{\sigma_{\max_{+/-}}^2}{\bar{T}} \right)
\end{aligned}$$

Remarkably, the dependency of n nearly vanishes if we again assume $\sigma_{\max_{+/-}}^2$ not to be affected by n. It is hidden only in $\bar{T} = T - 2n - 2$. This implies that *Stratified SVARM* has the hypothetical potential to exhibit approximation quality which almost does not deteriorate with growing player numbers, or the size of the cooperative game in other words. To the best of our knowledge, no other approximation algorithm for the Shapley value possesses this analytical property. For example, employing the Neyman allocation (see Equation 4.7) for stratified sampling of marginal contributions leads to an MSE of

$$\mathbb{E}[\text{MSE}] = \frac{2}{n^3 \bar{T}} \left(\sum_{i=1}^{n} \sum_{\ell=1}^{n-2} \sigma_{i,\ell} \right)^2 \in \mathcal{O}\left(\frac{\sigma_{\max}^2 \cdot n}{\bar{T}} \right) . \tag{4.23}$$

As already alluded to, the stratum variances are inaccessible to the approximation algorithm and have to be estimated. Thus, we transfer the approach of Castro et al. (2017) and propose *Adaptive SVARM* which divides its available budget into two phases conducting exploration and exploitation. The first phase collects samples from all strata to maintain estimates $\hat{\sigma}^2_{i,\ell,+}$ and $\hat{\sigma}^2_{i,\ell,-}$. Before entering the second phase, the presumably optimal allocation $\hat{m}^*_\ell$ is computed according to Equation 4.21 based on the stratum variance estimates. Within the second phase *Adaptive SVARM* does not allocate its samples as prescribed by $\hat{m}^*_\ell$ as this would only be optimal for the second phase in isolation. Instead, it fills up the total allocation including the first phase such that the stratification is optimal across the entire budget at disposal. This is done by pursuing in the second phase the allocation that results from the difference between the optimal allocation for $\bar{T}$ and that realized in the first phase. Consequently, it reaches $\hat{m}^*_\ell$ at termination. Moreover, we develop *Continuous Adaptive SVARM* as an extension that keeps estimating the stratum variances during the second phase and simultaneously adapts $\hat{m}_\ell$. The underlying idea is that the observations made in the second phase are equally valid to improve the precision of $\hat{\sigma}^2_{i,\ell,+}$ and $\hat{\sigma}^2_{i,\ell,-}$ and thus more reliably estimate m^*_ℓ.

Our empirical results convey two messages. First, depending on the cooperative game, the performance gap between *Stratified SVARM* and its optimal version in hindsight is indeed significant. Second, both of our adaptive algorithms close this gap and thus bring the hypothetical potential of stratification to life.

Contribution (III). We further assess the performance of *Stratified SVARM* and compare stratified sampling against methods representing other branches of approximation. In particular, cooperative games constructed to derive feature importance scores for unlabeled data are considered. In the absence of target variables and a predictive model, Balestra et al. (2022) propose the total correlation between discrete features $X_1, \ldots, X_n$ as a measure of worth for a coalition S of features, i.e.

$$\nu(S) = \left(\sum_{X_i \in S} H(X_i) \right) - H((X_i)_{i \in S}) \,, \tag{4.24}$$

where $H(\cdot)$ denotes the Shannon entropy. As a result, the retrieved Shapley values should indicate the features' usefulness in unsupervised learning. We interpret the obtained empirical findings as evidence for the effectiveness of stratification in general and in particular that of *Stratified SVARM*. Clearly, a dependence of a coalition's worth to its size reduces stratum variances and thus speeds up the

convergence of mean estimates. This is plausibly the case for the total correlation of features as it is likely to increase with coalition size.

4.3 Shapley Value Approximation via Optimization

Contribution (IV). Inspecting *KernelSHAP* (Lundberg and Lee, 2017) from a more conceptual standpoint, we recognize that it is not only a direct application of the Shapley value's representation as a solution to a weighted regression problem (Charnes et al., 1988), but rather an instance of a more general framework that we establish. As described in Section 4.1, a surrogate game is fitted to match the given value function ν and its own Shapley values are taken as estimates. The simplicity of the surrogate game utilized by *KernelSHAP* (see Equation 4.12) incentivizes us to plug in a more sophisticated value function ν'. Greater flexibility of ν' should ease to capture more intricate patterns of coalition values and therefore promise a better fit to ν. We take inspiration from Grabisch (1997a) stating that any value function can be decomposed into Shapley interactions as

$$\nu(S) = \sum_{A \subseteq \mathcal{N}} \gamma_{A,S} \cdot I_A^\phi \quad \text{for all } S \subseteq \mathcal{N} \tag{4.25}$$

with suitable coefficients $\gamma_{A,S}$ that do not depend on ν but only on the coalitions A and S themselves. Keeping the interaction values I_A^ϕ as free variables to adjust would allow to fit a value function ν' of that form to palpated points of ν. However, since each $A \subseteq \mathcal{N}$ possesses its own interaction, this representation exhibits 2^n degrees of freedom which prohibits a unique identifiable solution to the resulting optimization problem (similar to Equation 4.13) when only a subset of coalitions is evaluated. As a remedy, we truncate the surrogate game's flexibility by restricting interactions up to a chosen order k which yields the k-*additive* value function

$$\nu_k(S) = \sum_{\substack{A \subseteq \mathcal{N} \\ |A| \leq k}} \gamma_{A,S} \cdot I_A^k \quad \text{for all } S \subseteq \mathcal{N}\,. \tag{4.26}$$

The concept of k-additivity (Grabisch, 1997b) provides a useful tool to impose structure upon a value function, or discrete fuzzy measure in general, by demanding from it to be additive up to some order k. In our case this translates to setting all Shapley interactions of higher order than k to zero. This is not even far-fetched, in the context of machine learning it is typical for higher order interactions between features to diminish and being close to zero (Bordt and Luxburg, 2023), while other

empirical works demonstrate how 2-additive and 3-additive measures suffice for appropriate modeling (Grabisch et al., 2006; Pelegrina et al., 2020). Important to note is that also this surrogate game's representation immediately yields its own Shapley values in form of singleton interactions $I_i^k := I_{\{i\}}^k$.

Subsequently, our proposed method *SVAk*$_{\text{ADD}}$ samples coalitions $S_1, \ldots, S_T \in \mathcal{P}(\mathcal{N}) \setminus \{\emptyset, \mathcal{N}\}$ without replacement to fill the objective function of Equation 4.13 with ν' being substituted by ν_k and free variables I_A^k for all $A \subseteq \mathcal{N}$ with $|A| \leq k$:

$$\begin{aligned} \min_{I^k} \; & \sum_{t=1}^{T} w_{S_T} \left(\nu_k(S_t) - \nu(S_t)\right)^2 \\ \text{s.t.} \; & \sum_{i=1}^{n} I_i^k = \nu(\mathcal{N}) - \nu(\emptyset) \end{aligned} \tag{4.27}$$

Here, each I_i^k represents the estimate $\hat{\phi}_i$. Although ν_k is fitted to ν, one might rightfully ask whether the estimates converge to the desired Shapley values ϕ_i of $(\mathcal{N}, \nu)$. We clear these doubts by analytically showing that the solution to Equation 4.27 yields the Shapley values in the cases of $k = 1, 2, 3$ when the objective function is filled with all coalitions $S \in \mathcal{P}(\mathcal{N}) \setminus \{\emptyset, \mathcal{N}\}$ and again weights $w_S = \binom{n-2}{|S|-1}^{-1}$ are being used, i.e. $I_i^1 = I_i^2 = I_i^3 = \phi_i$ for all $i \in \mathcal{N}$. We interpret this result as evidence for the soundness of our approach and appropriateness of ν_k. Moreover, our result is oblivious to the shape of ν. The complete optimization problem yields exact Shapley values despite ν not even being proximately k-additive. Hence, its solution constitutes a novel representation of the Shapley value.

4.4 Approximation of Shapley Interactions

Contribution (V). The fact that the Shapley interaction index originates from the Shapley value, or more generally speaking the cardinal interactions can be viewed as descendants of semivalues, entices to project approximation methods from the Shapley value to Shapley interactions. Being aware of the shared similarities between both quantities, we extend *Stratified SVARM* from (I) to approximate any cardinal interaction index. This involves generalizing key ideas of *Stratified SVARM*, lifting them to higher order. Starting with the representation of the interaction I_K for any

cardinal interaction index (see Definition 2.21), we split the discrete derivatives $\Delta_K(S)$ into coalition values according to Proposition 2.20 and stratify by size:

$$I_K = \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} w_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \underbrace{\frac{1}{\binom{n-k}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S|=\ell}} \nu(S \cup W)}_{=: I_{K,\ell}^W}, \tag{4.28}$$

where $k = |K|$. The algebraic manipulation involving the factor $\binom{n-k}{\ell}$ allows us to form strata values $I_{K,\ell}^W$ that are an arithmetic mean of coalition values, suitable for mean estimation via sampling. Note that these strata form a unique partitioning of $\mathcal{P}(\mathcal{N})$ for each K because every coalition $S \cup W \subseteq N$, is contained in exactly one stratum. Any cardinal interaction index can be expressed by Equation 4.28 as plugging in its associated weights $w_{k,\ell}$ recovers Definition 2.21.

On this basis, we propose the approximation algorithm *SVARM-IQ* that samples coalitions and updates its estimates analogously to *Stratified SVARM*. Each evaluated worth $\nu(A)$ is used to update an estimate $\hat{I}_{K,\ell}^W$ with $W = A \cap K$ and $\ell = |A \setminus W|$ for each $K \subseteq \mathcal{N}$ of size k. For each K the stratum estimates are aggregated to

$$\hat{I}_K = \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} w_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \hat{I}_{K,\ell}^W \tag{4.29}$$

such that the maximum sample reuse principle is again fulfilled since every observed $\nu(A)$ contributes to all estimates. As the stratum estimates can be weighted to form any cardinal interaction, *SVARM-IQ* is capable of approximating an arbitrary selection of indices simultaneously if the orders of interest k are provided in advance. This includes the Shapley value and semivalues, extending *Stratified SVARM*. Moreover, the indices do not have to be specified a priori because the aggregation in Equation 4.29 can be performed on demand and after the completion of sampling.

A more extensive overview of interaction indices is given in (Muschalik et al., 2024) and the comparison to state-of-the-art methods speaks in favor of *SVARM-IQ* for many use cases appearing within the field of machine learning.

4.5 Top-k Shapley Players Identification

Contribution (VI). Given the many approximation algorithms to choose from, one could simply reduce the top-k identification problem to the approximate-all problem by estimating all players' Shapley values precisely and returning those players with the highest estimates. Narayanam and Narahari (2008) put this thought to practice, employing an algorithm equivalent to *ApproShapley* (Castro et al., 2009). However, this approach might be viewed as too naive since some player's Shapley values turn out to be so low during approximation, or at least their estimates, that these can be excluded from the top-k with high confidence. The same holds true for the other end of the spectrum inhabited by players with relatively high estimates. As a result of this observation, it is perspicuous to favor the sampling for players who seem to be close to the border between the top-k and the rest, as their membership is more difficult to assess. Hence, the precision of players' estimates clearly belonging to the top-k, or not, is sacrificed to save budget for critical players and speed up the convergence to the desired segregation.

This algorithmic idea has already been discovered by the field of multi-armed bandits, a subbranch of online learning, initially to identify the distribution with highest mean (Bubeck et al., 2009), corresponding to top-1, and later for the general case of top-k (Gabillon et al., 2011; Bubeck et al., 2013). Here, metaphorically, one considers n many so-called arms $a_1, \ldots, a_n$ of a slot machine, each of which yields a numerical reward upon pulling. The latent reward distributions are unknown and potentially mutually different, likewise, their mean rewards. Pulling an arm a_i corresponds to drawing a random sample r_i from its reward distribution.

To the best of our knowledge, we are first to introduce the connection between top-k Shapley players identification and multi-armed bandits by formulating each player $i \in \mathcal{N}$ as an arm a_i, and its distribution of marginal contributions w.r.t. to the weights within the Shapley value as the arm's reward distribution. Thus, by pulling an arm a_i, one observes a random marginal contribution $\Delta_i(S)$ of that player as a reward r_i. This facilitates the immediate application of multi-armed bandit algorithms such as *Gap-E* (Gabillon et al., 2011) and *SAR* (Bubeck et al., 2013).

Moreover, we propose *Border Uncertainty Sampling* (BUS), a bandit algorithm that greedily selects the next player which optimizes a selection criterion. The criterion intertwines exploration and exploitation by favoring players with low sample numbers and estimates on the verge of belonging to the top-k respectively.

Contribution (VII). Despite the usage of estimates $\hat{\phi}_i$, one for each player, it suffices to correctly predict the ordering of players w.r.t. their Shapley values without even knowing these values precisely. A step towards this direction can be made by pairwise comparing estimates and judging for each pair of players whether $\phi_i < \phi_j$ holds, approximated by $\hat{\phi}_i < \hat{\phi}_j$. This approach is resilient against distortions that impact both estimates equally. Hence, precise estimates are not necessarily required, but it is rather their difference that plays a role. Pursuing this idea, we analytically show that for algorithms relying on Monte Carlo estimates, it is advantageous to incorporate a positive covariance between $\hat{\phi}_i$ and $\hat{\phi}_j$ into their sampling procedure such that the risk of a misordering is reduced. The underlying variance reduction technique of leveraging covariance is known as *antithetic sampling*.

Inspired by this observation, we intend to sample marginal contributions for all players that share a coalition, foregoing the independence of the player's estimates. In particular, for any $S \subseteq \mathcal{N}$, the marginal contributions $\Delta_i(S)$ of all $i \in \mathcal{N} \setminus S$ and $\Delta_j(S \setminus \{j\})$ of all $j \in S$ share the worth $\nu(S)$. We cover both cases by the unifying notion of the *extended marginal contribution*

$$\Delta'_i(S) := \nu(S \cup \{i\}) - \nu(S \setminus \{i\})\,. \tag{4.30}$$

Intuitively, the collected observations should be positively correlated as extended marginal contributions of different players to the same coalition are more likely similar in value than independent draws. On this basis, we introduce *Comparable Marginal Contributions Sampling (CMCS)* which round-wise collects extended marginal contributions for all players to maintain estimates $\hat{\phi}_i$. To our favor, sampling extended marginal contributions does not introduce a bias when a distribution over $\mathcal{P}(\mathcal{N})$ is employed that corresponds to the weights in the following novel representation of the Shapley value:

$$\phi_i = \sum_{S \subseteq \mathcal{N}} \frac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta'_i(S)\,. \tag{4.31}$$

In addition, *CMCS* displays a certain budget-efficiency by only requiring $n+1$ evaluations for n many updates in comparison to the independent sampling of marginal contributions which consumes two per update. However, it performs pure exploration as it is not selective about the players that it samples for. Therefore, our extension *Greedy CMCS* tackles the exploration-exploitation dilemma and saves budget by randomly leaving out players whose membership is relatively certain. Adopting the top-k algorithm of Kariyappa et al. (2024), our variant *CMCS@k* greedily selects in each round only two players guided by their overlap in confidence intervals.

5 Approximating the Shapley Value without Marginal Contributions

Author Contribution Statement

The author alone developed the idea and algorithms. The author wrote the analysis and experiment design with editing and proofreading by Viktor Bengs. The author conducted all experiments, and Patrick Becker implemented most algorithms and the simulation environment following the author's instructions. Maximilian Muschalik contributed by implementing experiments to test on real-world models and datasets. The author created the visualizations. The author has written the paper with the support of Viktor Bengs and proofreading by Eyke Hüllermeier.

Supplementary Material

An appendix to the paper is provided in Appendix A.

Approximating the Shapley Value without Marginal Contributions

Patrick Kolpaczki[1], **Viktor Bengs**[2,3], **Maximilian Muschalik**[2,3], **Eyke Hüllermeier**[2,3]

[1]Paderborn University
[2]Institute of Informatics, University of Munich (LMU)
[3]Munich Center for Machine Learning
patrick.kolpaczki@upb.de, viktor.bengs@lmu.de, maximilian.muschalik@lmu.de, eyke@lmu.de

Abstract

The Shapley value, which is arguably the most popular approach for assigning a meaningful contribution value to players in a cooperative game, has recently been used intensively in explainable artificial intelligence. Its meaningfulness is due to axiomatic properties that only the Shapley value satisfies, which, however, comes at the expense of an exact computation growing exponentially with the number of agents. Accordingly, a number of works are devoted to the efficient approximation of the Shapley value, most of them revolve around the notion of an agent's marginal contribution. In this paper, we propose with *SVARM* and *Stratified SVARM* two parameter-free and domain-independent approximation algorithms based on a representation of the Shapley value detached from the notion of marginal contribution. We prove unmatched theoretical guarantees regarding their approximation quality and provide empirical results including synthetic games as well as common explainability use cases comparing ourselves with state-of-the-art methods.

Introduction

Whenever agents can federalize in groups (form coalitions) to accomplish a task and get rewarded with a collective benefit that is to be shared among the group members, the notion of *cooperative game* stemming from game theory is arguably the most favorable concept to model such situations. This is due to its simplicity, which nevertheless allows for covering a whole range of practical applications. The agents are called *players* and are contained in a player set $\mathcal{N}$. Each possible subset of players $S \subseteq \mathcal{N}$ is understood as a *coalition* and the coalition $\mathcal{N}$ containing all players is called the *grand coalition*. The collective benefit $\nu(S)$ that a coalition S receives upon formation is given by a *value function* ν assigning each coalition a real-valued *worth*.

The connection of cooperative games to (supervised) machine learning is already well-established. The most prominent example is feature importance scores, both local and global, for a machine learning model: features of a dataset can be seen as players, allowing one to interpret a feature subset as a coalition, while the model's generalization performance using exactly that feature subset is its worth (Cohen, Dror, and Ruppin 2007). Other applications include evaluating the importance of parameters in a machine learning model, e.g. single neurons in a deep neural network (Ghorbani and Zou 2020) or base learners in an ensemble (Rozemberczki and Sarkar 2021), or assigning relevance scores to datapoints in a given dataset (Ghorbani and Zou 2019). See Rozemberczki et al. (2022) for a wider overview of its usage in the field of explainable artificial intelligence. Outside the realm of machine learning cooperative games also found applications in operations research (Luo, Zhou, and Lev 2022), for finding fair compensation mechanisms in electricity grids (O'Brien, Gamal, and Rajagopal 2015), or even for the purpose of identifying the most influential individuals in terrorist networks (van Campen et al. 2018).

In all of these applications, the question naturally arises of how to appropriately determine the contribution of a single player (feature, parameter, etc.) with respect to the grand collective benefit. In other words, how to allocate the worth $\nu(\mathcal{N})$ of the full player set $\mathcal{N}$ among the players in a fair manner. The indisputably most popular solution to this problem is the *Shapley value* (Shapley 1953), which can be intuitively expressed by *marginal contributions*. We call the increase in worth that comes with the inclusion of player i to a coalition S, i.e., the difference $\nu(S \cup \{i\}) - \nu(S)$, player i's marginal contribution to S. The Shapley value of i is a weighted average of all its marginal contributions to coalitions that do not include i. Its popularity stems from the fact that it is the only solution to satisfy axiomatic properties that arguably capture fairness (Shapley 1953).

Despite the appealing theoretical properties of the Shapley value, there is one major drawback with respect to its practical application, as its computational complexity increases exponentially with the number of players n. As a consequence, the exact computation of the Shapley value becomes practically infeasible even for a moderate number of players. This is especially the case where accesses to ν are costly, e.g., re-evaluating a (complex) machine learning model for a specific feature subset, or manipulating training data each time ν is accessed. Recently, several approximation methods have been proposed in search of a remedy, enabling the utilization of the Shapley value in explainable AI (and beyond). However, most works are stiffened towards the notion of marginal contribution, and, consequently, judge algorithms by their achieved approximation accuracy depending on the number of evaluated marginal

contributions. This measure does not do justice to the fact that approximations can completely dispense with the consideration of marginal contributions and elicit information from ν in a more efficient way — as we show in this paper. We claim that the number of single accesses to ν should be considered instead, since especially in machine learning, as mentioned above, access to ν is a bottleneck in overall runtime. In this paper, we make up for this deficit by considering the problem of approximating the Shapley values under a fixed *budget* T of evaluations (accesses) of ν.

Contribution. We present a novel representation of the Shapley value that does not rely on the notion of marginal contribution. Our first proposed approximation algorithm *Shapley Value Approximation without Requesting Marginals* (SVARM) exploits this representation and directly samples values of coalitions, facilitating "a swarm of updates", i.e., multiple Shapley value estimates are updated at once. This is in stark contrast to the usual way of sampling marginal contributions that only allows the update of a single estimate. We prove theoretical guarantees regarding SVARM's precision including the bound of $\mathcal{O}(\frac{\log n}{T-n})$ on its variance.

Based on a partitioning of the set of all coalitions according to their size, we develop with *Stratified SVARM* a refinement of SVARM. The applied stratification materializes a twofold improvement: (i) the homogeneous strata (w.r.t. the coalition worth) significantly accelerate convergence of estimates, (ii) our stratified representation of the Shapley value with decomposed marginal contributions facilitates a mechanism that updates the estimates of *all* players with *each single* coalition sampled. Among other results, we bound its variance by $\mathcal{O}\big(\frac{\log n}{T-n\log n}\big)$.

Besides our superior theoretical findings, both algorithms possess a number of properties in their favor. More specifically, both are unbiased, parameter-free, incremental, i.e., the available budget has not to be fixed and can be enlarged or cut prematurely, facilitating on-the-fly approximations due to their anytime property, and do not require any knowledge about the latent value function. Moreover, both are domain-independent and not limited to some specific fields, but can be used to approximate the Shapley values of any possible cooperative game.

Finally, we compare our algorithms empirically against other popular competitors, demonstrating their practical usefulness and proving our empirical enhancement *Stratified SVARM*$^+$, which samples without replacement to be the first sample-mean-based approach to achieve rivaling state-of-the-art approximation quality. All code including documentation and the technical appendix can be found on GitHub[1].

Related Work

The recent rise of explainable AI has incentivized the research on approximation methods for the Shapley value leading to a variety of different algorithms for this purpose. The first distinction to be made is between those that are domain-independent, i.e., able to deal with any cooperative game, and those that are tailored to a specific use case, e.g. assigning Shapley values to single neurons in neural networks, or which impose specific assumptions on the value function. In this paper, we will consider only the former, as it is our goal to provide approximations algorithms independent of the context in which they are applied. The first and so far simplest of this kind is *ApproShapley* (Castro, Gómez, and Tejada 2009), which samples marginal contributions from each player based on randomly drawn permutations of the player set. The variance of each of its Shapley value estimates is bounded by $\mathcal{O}(\frac{n}{T})$. *Stratified Sampling* (Maleki et al. 2013) and *Structured Sampling* (van Campen et al. 2018) both partition the marginal contributions of each player by coalition size in order to stratify the marginal contributions of the population from which to draw a sample, which leads to a variance reduction. While *Stratified Sampling* calculates a sophisticated allocation of samples for each coalition size, *Structured Sampling* simply samples with equal frequencies. Multiple follow-up works suggest specific techniques to improve the sampling allocation over the different coalition sizes (O'Brien, Gamal, and Rajagopal 2015; Castro et al. 2017; Burgess and Chapman 2021).

In order to reduce the variance of the naive sampling approach underlying *ApproShapley*, Illés and Kerényi (2019) suggest to use ergodic sampling, i.e., generating samples that are not independent but still satisfy the strong Law of Large numbers. Quite recently, Mitchell et al. (2022) investigated two techniques for improving *ApproShapley*'s sampling approach. One is based on the theory of reproducing kernel Hilbert spaces, which focuses on minimizing the discrepancies for functions of permutations. The other exploits a geometrical connection between uniform sampling on the Euclidean sphere and uniform sampling over permutations.

Adopting a Bayesian perspective, i.e., by viewing the Shapley values as random variables, Touati, Radjef, and Sais (2021) consider approximating the Shapley values by Bayesian estimates (posterior mean, mode, or median), where each posterior distribution of a player's Shapley value depends on the remaining ones. Utilizing a representation of the Shapley value as an integral (Owen 1972), *Owen Sampling* (Okhrati and Lipani 2020) approximates this integral by sampling marginal contributions using antithetic sampling (Rubinstein and Kroese 2016; Lomeli et al. 2019) for variance reduction.

A fairly new class of approaches that dissociates itself from the notion of marginal contribution are those that view the Shapley value as a solution of a quadratic program with equality constraints (Lundberg and Lee 2017; Simon and Vincent 2020; Covert and Lee 2021). Another unorthodox approach is to divide the player set into small enough groups for which the Shapley values within these groups can be computed exactly (Soufiani et al. 2014; Corder and Decker 2019). For an overview of approaches related to machine learning we refer to (Chen et al. 2023).

Problem Statement

The formal notion of a cooperative game is defined by a tuple $(\mathcal{N}, \nu)$ consisting of a set of players $\mathcal{N} = \{1, \dots, n\}$ and a value function $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ that assigns to each subset

[1] https://github.com//kolpaczki//Approximating-the-Shapley-Value-without-Marginal-Contributions

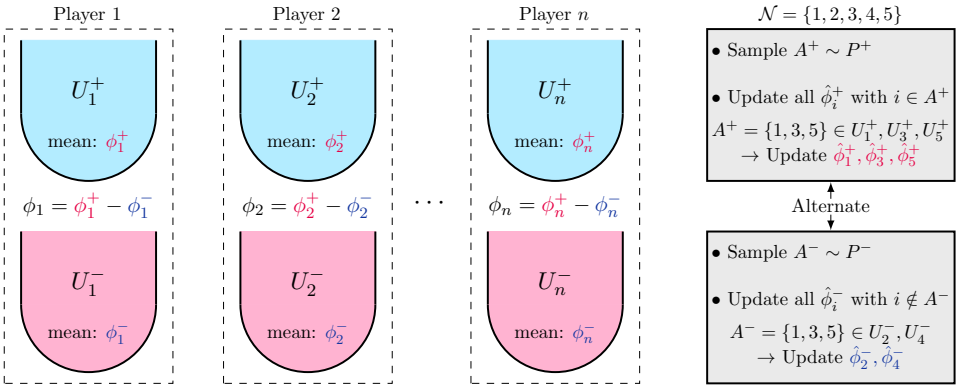


Figure 1: Illustration of SVARM's sampling process and update rule: Each player i has two urns $U_i^+ := \{S \cup \{i\} \mid S \subseteq \mathcal{N}_i\}$ and $U_i^- := \{S \mid S \subseteq \mathcal{N}_i\}$ containing marbles which represent coalitions, with mean coalition worth ϕ_i^+ and ϕ_i^-. SVARM alternates between sampling coalitions $A^+ \sim P^+$ and $A^- \sim P^-$. With each drawn coalition all estimates of those urns are updated which contain the corresponding marble. Since each player's two urns form a partition of the powerset $\mathcal{P}(\mathcal{N})$, all players have exactly one urn updated with each sample.

of $\mathcal{N}$ a real-valued number. The value function must satisfy $\nu(\emptyset) = 0$. We call the subsets of $\mathcal{N}$ coalitions, $\mathcal{N}$ itself the grand coalition, and the assigned value $\nu(S)$ to a coalition $S \subseteq \mathcal{N}$ its worth. Given a cooperative game $(\mathcal{N}, \nu)$, the Shapley value assigns each player a share of the grand coalition's worth. In particular, the Shapley value (Shapley 1953) of any player $i \in \mathcal{N}$ is defined as

$$\phi_i = \sum_{S \subseteq \mathcal{N}_i} \frac{1}{n \cdot \binom{n-1}{|S|}} \left[\nu(S \cup \{i\}) - \nu(S)\right], \tag{1}$$

where $\mathcal{N}_i := \mathcal{N} \setminus \{i\}$ for each player $i \in \mathcal{N}$. The term $\nu(S \cup \{i\}) - \nu(S)$ is also known as player i's marginal contribution to $S \subseteq \mathcal{N}_i$ and captures the increase in collective benefit when player i joins the coalition S. Thus, the Shapley value can be seen as the weighted average of a player's marginal contributions.

The exact computation of all Shapley values requires the knowledge of the values of all 2^n many coalitions[2] and is shown to be NP-hard (Deng and Papadimitriou 1994). In light of the exponential computational effort w.r.t. to n, we consider the goal of approximating the Shapley value of all players as precisely as possible for a given *budget* of $T \in \mathbb{N}$ many evaluations (accesses) of ν in discrete time steps $1, \ldots, T$. Since $\nu(\emptyset) = 0$ holds by definition, the evaluation of $\nu(\emptyset)$ comes for free without any budget cost. We judge the quality of the estimates $\hat{\phi}_1, \ldots, \hat{\phi}_n$ — which are possibly of stochastic nature — obtained by an approximation algorithm after T many evaluations by two criteria that have to be minimized for all $i \in \mathcal{N}$. First, the mean squared error (MSE) of the estimate $\hat{\phi}_i$ is given by

$$\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right]. \tag{2}$$

Utilizing the bias-variance decomposition allows us to reduce the squared error to the variance $\mathbb{V}[\hat{\phi}_i]$ of the Shapley value estimate in case that it is unbiased, i.e. $\mathbb{E}[\hat{\phi}_i] = \phi_i$. The second criterion is the probability of $\hat{\phi}_i$ deviating from ϕ_i by more than a fixed $\varepsilon > 0$:

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| > \varepsilon\right). \tag{3}$$

Both criteria are well-established for measuring the quality of an algorithm approximating the Shapley value.

SVARM

Thanks to the distributive law, the formula of the Shapley value for a player i can be rearranged so that it is not its weighted average of marginal contributions, but the difference of the weighted average of coalition values by adding i and the weighted average of coalition values without i:

$$\phi_i = \underbrace{\sum_{S \subseteq \mathcal{N}_i} w_S \cdot \nu(S \cup \{i\})}_{=: \phi_i^+} - \underbrace{\sum_{S \subseteq \mathcal{N}_i} w_S \cdot \nu(S)}_{=: \phi_i^-}, \tag{4}$$

with weights $w_S = \frac{1}{n \cdot \binom{n-1}{|S|}}$ for each $S \subseteq \mathcal{N}_i$. We call ϕ_i^+ the positive and ϕ_i^- the negative Shapley value, while we refer to the collective of both as the signed Shapley values. The weighted averages ϕ_i^+ and ϕ_i^- can also be viewed as expected values, i.e., $\phi_i^+ = \mathbb{E}[\nu(\mathcal{S} \cup \{i\})]$ and $\phi_i^- = \mathbb{E}[\nu(\mathcal{S})]$, where $\mathcal{S} \sim P^w$ and $P^w(S) = w_S$ for all $S \subseteq \mathcal{N}_i$. Note that all weights add up to 1 and thus P^w forms a well-defined probability distribution. In this way, we can approximate each signed Shapley value separately using estimates $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ and combine them into a Shapley value estimate by means of $\hat{\phi}_i = \hat{\phi}_i^+ - \hat{\phi}_i^-$.

In light of this, a naive approach for approximating each signed Shapley value of a player is by sampling some number of M many coalitions $S^{(1)}, \ldots, S^{(M)}$ with distribution P^w and using the sample mean as the estimate, i.e., $\hat{\phi}_i^+ = \frac{1}{M} \sum_{m=1}^{M} \nu(S^{(m)} \cup \{i\})$. However, this would require all $2n$ signed Shapley values (two per player) to be

[2] In fact, only $2^n - 1$ many coalitions, as $\nu(\emptyset) = 0$ is known.

estimated separately by sampling coalitions in a dedicated manner, each of which would lead to an update of only one estimate. This ultimately slows down the convergence of the estimates, especially for large n.

On the basis of the aforementioned representation of the Shapley value, we present the *Shapley Value Approximation without Requesting Marginals* (SVARM) algorithm, a novel approach that updates multiple Shapley value estimates at once with a single evaluation of ν. Its novelty consists of sampling coalitions independently from two specifically chosen distributions P^+ and P^- in an alternating fashion, which allows for a more powerful update rule: each (independently) sampled coalition A^+ from P^+ allows one to update all positive Shapley value estimates $\hat{\phi}_i^+$ of all payers i which are contained in A^+, i.e., $i \in A^+$. Likewise, for a coalition A^- drawn from P^-, all negative Shapley value estimates $\hat{\phi}_i^-$ for $i \notin A^-$ can be updated.

It is worth noting that, for simplicity, we alternate evenly between the samples from the P^+ and P^- distributions, although one could also use a ratio other than $1/2$. To avoid a bias, both distributions have to be tailored such that the following holds for all $i \in \mathcal{N}$ and $S \subseteq \mathcal{N}_i$:

$$\begin{aligned}\mathbb{P}(A^+ = S \cup \{i\} \mid i \in A^+) = \\ = \mathbb{P}(A^- = S \mid i \notin A^-) = w_S\,.\end{aligned} \tag{5}$$

For this reason, we define the probability distributions over coalitions to sample from as

$$P^+(S) := \frac{1}{|S| \binom{n}{|S|} H_n} \qquad \forall S \in \mathcal{P}(\mathcal{N}) \setminus \{\emptyset\}, \tag{6}$$

$$P^-(S) := \frac{1}{(n - |S|) \binom{n}{|S|} H_n} \qquad \forall S \in \mathcal{P}(\mathcal{N}) \setminus \{\mathcal{N}\}, \tag{7}$$

where $H_n = \sum_{k=1}^{n} 1/k$ denotes the n-th harmonic number. Note that both P^+ and P^- assign equal probabilities to coalitions of the same size, so that one can first sample the size and then draw a set uniformly of that size. This pair of distributions is provably the only one to fulfill the required property (see Appendix C.1).

The approach of dividing the Shapley value into two parts and approximating both has already been pursued (although not as formally rigorous) via importance sampling (Covert, Lundberg, and Lee 2019), allowing to update all n estimates with each sample. Wang and Jia (2023) adopt the same representation for the Banzhaf value, and coined the strategy of updating all players' estimates with each sampled coalition the *maximum sample reuse* (MSR) principle. Their approximation algorithm is specifically tailored to the Banzhaf value as it leverages its uniform weights $w_S = \frac{1}{2^{n-1}}$ and is thus, at least not directly, transferable to the Shapley value.

In the following we describe SVARM's procedure with the pseudocode of Algorithm 1. The overall idea of the sampling and update process is illustrated in Figure 1. It starts by initializing the positive and negative Shapley value estimates $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$, and the number of samples c_i^+ and c_i^- collected for each player i. SVARM continues by launching a warm-up phase (see Algorithm 3 in Appendix B). In the main loop, the update rule is applied for as many sampled pairs of coalitions A^+ and A^- as possible until SVARM runs out of budget. In each iteration A^+ is sampled from P^+ and A^- from P^-. The worth of A^+ and A^- is evaluated and stored in v^+ and v^-, requiring two accesses to the value function. The estimate $\hat{\phi}_i^+$ of each player $i \in A^+$ is updated with the worth $\nu(A^+)$ such that $\hat{\phi}_i^+$ is the mean of sampled coalition values. Likewise, the estimate $\hat{\phi}_i^-$ of each player $i \notin A^-$ is updated with the worth $\nu(A^-)$. At the same time, the sample numbers of the respective signed Shapley value estimates are also updated. Finally, SVARM computes its Shapley value estimate $\hat{\phi}_i$ of ϕ_i for each i according to Equation (4). Note that since only the quantities $\hat{\phi}_i^+, \hat{\phi}_i^-, c_i^+$, and c_i^+ are stored for each player, its space complexity is in $\mathcal{O}(n)$. Moreover, SVARM is incremental and can be stopped at any time to return its estimates after executing line 20, or it can be run further with increased budget.

Algorithm 1: SVARM

Input: $\mathcal{N}, T \in \mathbb{N}$

1: $\hat{\phi}_i^+, \hat{\phi}_i^- \leftarrow 0$ for all $i \in \mathcal{N}$
2: $c_i^+, c_i^- \leftarrow 1$ for all $i \in \mathcal{N}$
3: WarmUp
4: $t \leftarrow 2n$
5: **while** $t + 2 \leq T$ **do**
6: Draw $A^+ \sim P^+$
7: Draw $A^- \sim P^-$
8: $v^+ \leftarrow \nu(A^+)$
9: $v^- \leftarrow \nu(A^-)$
10: **for** $i \in A^+$ **do**
11: $\hat{\phi}_i^+ \leftarrow \frac{c_i^+ \hat{\phi}_i^+ + v^+}{c_i^+ + 1}$
12: $c_i^+ \leftarrow c_i^+ + 1$
13: **end for**
14: **for** $i \in \mathcal{N} \setminus A^-$ **do**
15: $\hat{\phi}_i^- \leftarrow \frac{c_i^- \hat{\phi}_i^- + v^-}{c_i^- + 1}$
16: $c_i^- \leftarrow c_i^- + 1$
17: **end for**
18: $t \leftarrow t + 2$
19: **end while**
20: $\hat{\phi}_i \leftarrow \hat{\phi}_i^+ - \hat{\phi}_i^-$ for all $i \in \mathcal{N}$

Output: $\hat{\phi}_1, \ldots, \hat{\phi}_n$

Theoretical Analysis. In the following we present theoretical results for SVARM. All proofs are given in Section C of the technical appendix. For the remainder of this section we assume that a minimum budget of $T \geq 2n + 2$ is given. This assumption guarantees the completion of the warm-up phase such that each positive and negative Shapley value estimate has at least one sample and an additional pair sampled in the loop. The lower bound on T is essentially twice the number of players n, which is a fairly weak assumption. We denote by $\bar{T} := T - 2n$ the number of time steps (budget) left after the warm-up phase. Moreover, we assume $\bar{T}$ to be even for sake of simplicity such that a lower bound on the number of sampled pairs in the main part can be expressed

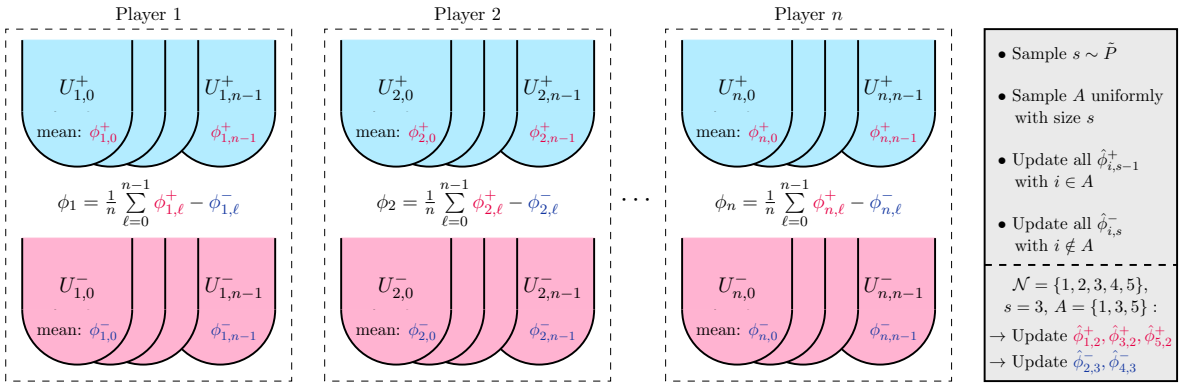


Figure 2: Illustration of Stratified SVARM's sampling process and update rule: Each player i has urns $U^+_{i,\ell} := \{S \cup \{i\} \mid S \subseteq \mathcal{N}_i, |S| = \ell\}$ and $U^-_{i,\ell} := \{S \mid S \subseteq \mathcal{N}_i, |S| = \ell\}$ for all $\ell \in \{0, \ldots, n-1\}$, $2n$ in total, containing marbles which represent coalitions, with mean coalition worth $\phi^+_{i,\ell}$ and $\phi^-_{i,\ell}$. Stratified SVARM samples in each time step t a coalition $A_t \subseteq \mathcal{N}$ and updates the estimates of all players' urns that contain the corresponding marble. Since each player's urns form a partition of the powerset $\mathcal{P}(\mathcal{N})$, all players have exactly one urn updated with each sample.

by $\frac{T}{2} - n$. We begin with the unbiasedness of the estimates maintained by SVARM allowing us later to reduce the mean squared error (MSE) of each estimate to its variance.

Theorem 1. *The Shapley value estimate $\hat{\phi}_i$ of any $i \in \mathcal{N}$ obtained by SVARM is unbiased, i.e.,*

$$\mathbb{E}[\hat{\phi}_i] = \phi_i.$$

Next, we give a bound on the variance of each Shapley value estimate. For this purpose, we introduce notation for the variances of coalition values contained in ϕ^+_i and ϕ^-_i. For a random set $A_i \subseteq \mathcal{N}_i$ distributed according to P^w let

$$\sigma_i^{+2} := \mathbb{V}\left[\nu(A_i \cup \{i\})\right] \text{ and } \sigma_i^{-2} := \mathbb{V}\left[\nu(A_i)\right]. \tag{8}$$

Theorem 2. *The variance of any player's Shapley value estimate $\hat{\phi}_i$ obtained by SVARM is bounded by*

$$\mathbb{V}[\hat{\phi}_i] \leq \frac{2H_n}{\bar{T}}(\sigma_i^{+2} + \sigma_i^{-2}).$$

Combining the unbiasedness in Theorem 1 with the latter variance bound implies the following result on the MSE.

Corollary 1. *The MSE of any player's Shapley value estimate $\hat{\phi}_i$ obtained by SVARM is bounded by*

$$\mathbb{E}\left[(\hat{\phi}_i - \phi_i)^2\right] \leq \frac{2H_n}{\bar{T}}(\sigma_i^{+2} + \sigma_i^{-2}).$$

Assuming that each variance term σ_i^{+2} and σ_i^{-2} is bounded by some constant independent of n (and T), the MSE bound in Corollary 1 is in $\mathcal{O}(\frac{\log n}{T-n})$ and so is the variance bound in Theorem 2. Note that this assumption is rather mild and satisfied if the underlying value function is bounded by constants independent of n, which again is the case for a wide range of games and in particular in explainable AI for global and local feature importance based on classification probabilities lying between 0 and 1. Further, as T is growing linearly with n by assumption, the denominator is essentially driven by the asymptotics of T. Thus, the dependency on n is logarithmic, which is a significant improvement over existing theoretical results having a linear dependency on n like $\mathcal{O}(\frac{n}{T})$ for *ApproShapley* (Castro, Gómez, and Tejada 2009) or possibly worse (Simon and Vincent 2020). Finally, we present two probabilistic bounds on the approximated Shapley value. The first utilizes the variance bound shown in Theorem 2 by applying Chebyshev's inequality.

Theorem 3. *The probability that the Shapley value estimate $\hat{\phi}_i$ of any fixed player $i \in \mathcal{N}$ deviates from ϕ_i by a margin of any fixed $\varepsilon > 0$ or greater is bounded by*

$$\mathbb{P}(|\hat{\phi}_i - \phi_i| \geq \varepsilon) \leq \frac{2H_n}{\varepsilon^2 \bar{T}}(\sigma_i^{-2} + \sigma_i^{+2}).$$

The presented bound is in $\mathcal{O}(\frac{\log n}{T-n})$ and improves upon the bound derived by Chebyshev's inequality of $\mathcal{O}(\frac{n}{T})$ for *ApproShapley* (Maleki et al. 2013). Our second bound derived by Hoeffding's inequality is tighter, but requires the introduction of notation for the ranges of $\nu(A_i)$ and $\nu(A_i \cup \{i\})$:

$$r_i^+ := \max_{S \subseteq \mathcal{N}_i} \nu(S \cup \{i\}) - \min_{S \subseteq \mathcal{N}_i} \nu(S \cup \{i\}), \tag{9}$$

$$r_i^- := \max_{S \subseteq \mathcal{N}_i} \nu(S) - \min_{S \subseteq \mathcal{N}_i} \nu(S). \tag{10}$$

Theorem 4. *The probability that the Shapley value estimate $\hat{\phi}_i$ of any fixed player $i \in \mathcal{N}$ deviates from ϕ_i by a margin of any fixed $\varepsilon > 0$ or greater is bounded by*

$$\mathbb{P}(|\hat{\phi}_i - \phi_i| \geq \varepsilon) \leq 2e^{-\frac{\bar{T}}{4H_n^2}} + 4\frac{e^{-\Psi\left\lfloor\frac{\bar{T}}{4H_n}\right\rfloor}}{e^{\Psi} - 1},$$

where $\Psi = 2\varepsilon^2/(r_i^+ + r_i^-)^2$.

Note that this bound is exponentially decreasing with T and can be expressed asymptotically as $\mathcal{O}(e^{-\frac{T-n}{(\log n)^2}})$. In comparison, the bounds of $\mathcal{O}(e^{-\frac{T}{n}})$ for *ApproShapley*, $\mathcal{O}(ne^{-\frac{T}{n^3}})$ for *Stratified Sampling* (Maleki et al. 2013), and the projected SGD variant (Simon and Vincent 2020) show worse asymptotic dependencies on n in comparison.

Stratified SVARM

On the basis of the representation of the Shapley value in Equation (4), we develop another approximation algorithm named *Stratified SVARM* to further pursue and reach the maximum sample reuse principle. Its crux is a refinement of *SVARM* obtained by stratifying the positive and the negative Shapley value ϕ_i^+ and ϕ_i^-. We exploit the latter to develop an even more powerful update rule that allows for updating all players simultaneously with each single coalition sampled. Both, ϕ_i^+ and ϕ_i^- can be rewritten using stratification such that each becomes an average of strata, whereas the strata themselves are averages of the coalitions' worth:

$$\phi_i^+ = \frac{1}{n}\sum_{\ell=0}^{n-1} \frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S\subseteq\mathcal{N}_i \\ |S|=\ell}} \nu(S\cup\{i\}) =: \frac{1}{n}\sum_{\ell=0}^{n-1}\phi_{i,\ell}^+, \quad (11)$$

$$\phi_i^- = \frac{1}{n}\sum_{\ell=0}^{n-1} \frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S\subseteq\mathcal{N}_i \\ |S|=\ell}} \nu(S) \qquad =: \frac{1}{n}\sum_{\ell=0}^{n-1}\phi_{i,\ell}^-. \quad (12)$$

We call $\phi_{i,\ell}^+$ the ℓ-th positive Shapley subvalue and $\phi_{i,\ell}^-$ the ℓ-th negative Shapley subvalue for all $\ell \in \mathcal{L} := \{0,\dots,n-1\}$. Now, we can write ϕ_i as

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

Note that this representation of ϕ_i coincides with Equation 6 in (Ancona, Öztireli, and Gross 2019). Intuitively speaking at the example of ϕ_i^+ (and analogously for ϕ_i^-), we partition the population of coalitions contained in ϕ_i^+ into n strata. Each *stratum* $\phi_{i,\ell}^+$ comprises all coalitions which include the player i and have cardinality $\ell+1$. Instead of sampling directly for ϕ_i^+, the stratification allows one to sample coalitions from each stratum, obtain mean estimates $\hat{\phi}_{i,\ell}^+$, and aggregate them to

$$\hat{\phi}_i^+ = \frac{1}{n}\sum_{\ell=0}^{n-1}\hat{\phi}_{i,\ell}^+ \quad (14)$$

in order to obtain an estimate for ϕ_i^+. Due to the increase in homogeneity of the strata in comparison to their origin population, caused by the shared size and inclusion or exclusion of i for coalitions in the same stratum, one would expect the strata to have significantly lower variances and ranges resulting in approximations of better quality compared to SVARM. In combination with our bounds shown in Theorem 2 and Theorem 4, this should result in approximations of better quality. In the following we present further techniques for improvement which we apply for Stratified SVARM (Algorithm 2).

Exact Calculation. First, we observe that some strata contain very few coalitions. Thus, we calculate $\phi_{i,0}^+, \phi_{i,n-2}^+, \phi_{i,n-1}^+, \phi_{i,1}^-$, and $\phi_{i,n-1}^-$ for all players exactly by evaluating ν for all coalitions of size $1, n-1$, and n. This requires $2n+1$ many evaluations of ν (see Algorithm 5 in Appendix B). We already obtain $\phi_{i,0}^- = \nu(\emptyset) = 0$ by definition. As a consequence, we can exclude the sizes $0, 1, n-1$, and n from further consideration. We assume for the remainder that $n \geq 4$, otherwise we would have already calculated all Shapley values exactly.

Algorithm 2: Stratified SVARM

Input: $\mathcal{N}, T \in \mathbb{N}$

$\hat{\phi}_{i,\ell}^+, \hat{\phi}_{i,\ell}^- \leftarrow 0$ for all $i \in \mathcal{N}$ and $\ell \in \mathcal{L}$
$c_{i,\ell}^+, c_{i,\ell}^- \leftarrow 0$ for all $i \in \mathcal{N}$ and $\ell \in \mathcal{L}$
ExactCalculation($\mathcal{N}$)
WarmUp$^+$($\mathcal{N}$)
WarmUp$^-$($\mathcal{N}$)
$t \leftarrow 2n + 1 + 2\sum_{s=2}^{n-2}\lceil \frac{n}{s} \rceil$
while $t < T$ **do**
  Draw $s_t \sim \tilde{P}$
  Draw A_t from $\{S \subseteq \mathcal{N} \mid |S| = s_t\}$ uniformly
  Update(A_t)
  $t \leftarrow t+1$
end while
$\hat{\phi}_i \leftarrow \frac{1}{n}\sum_{\ell=0}^{n-1}\hat{\phi}_{i,\ell}^+ - \hat{\phi}_{i,\ell}^-$ for all $i \in \mathcal{N}$

Output: $\hat{\phi}_1, \dots, \hat{\phi}_n$

Refined Warm-Up. Next, we split the warm-up into two parts, one for the positive, the other for the negative Shapley subvalues (see Algorithm 6 and 7 in Appendix B). Each collects for each estimate $\hat{\phi}_{i,\ell}^+$ or $\hat{\phi}_{i,\ell}^-$, respectively, one sample and consumes a budget of $\sum_{s=2}^{n-2}\lceil \frac{n}{s} \rceil$.

Enhanced Update Rule. Thanks to the stratified representation of the Shapley value, we can enhance SVARM's update rule and update with each sampled coalition $A_t \subseteq \mathcal{N}$ the estimates $\hat{\phi}_{i,|A_t|-1}^+$ for all $i \in A_t$ and $\hat{\phi}_{i,|A_t|}^-$ for all $i \notin A_t$. Thus, we can update all estimates $\hat{\phi}_i$ at once with a single sample. This enhanced update step is given in Algorithm 4 (see Appendix B) and illustrated in Figure 2. In order to obtain unbiased estimates, it suffices to select an arbitrary size s of the coalition A to be sampled and draw A uniformly at random from the set of coalitions with size s. We go one step further and choose not only the coalition A, but also the size s randomly according to a specifically tailored probability distribution $\tilde{P}$ over $\{2,\dots,n-2\}$, which leads to simpler bounds in our theoretical analysis in which each stratum receives the same weight. We define for n even:

$$\tilde{P}(s) := \begin{cases} \frac{n\log n - 1}{2sn\log n\left(H_{\frac{n}{2}-1}-1\right)} & \text{if } s \leq \frac{n-2}{2} \\ \frac{1}{n\log n} & \text{if } s = \frac{n}{2} \\ \frac{n\log n - 1}{2(n-s)n\log n\left(H_{\frac{n}{2}-1}-1\right)} & \text{otherwise} \end{cases},$$

and for n odd: $\tilde{P}(s) := \begin{cases} \frac{1}{2s\left(H_{\frac{n-1}{2}}-1\right)} & \text{if } s \leq \frac{n-1}{2} \\ \frac{1}{2(n-s)\left(H_{\frac{n-1}{2}}-1\right)} & \text{otherwise} \end{cases}$.

Note that Stratified SVARM is incremental just as SVARM, but in contrast, requires quadratic space $\mathcal{O}(n^2)$ as it stores estimates and counters for each player *and* stratum.

Theoretical Analysis. Similar to SVARM, we present in the following our theoretical results for Stratified SVARM. All proofs are given in Appendix D. Again, we assume a minimum budget of $T \geq 2n + 1 + 2\sum_{s=2}^{n-2} \left\lceil \frac{n}{s} \right\rceil =: W \in \mathcal{O}(n \log n)$, guaranteeing the completion of the warm-up phase, and denote by $\bar{T} = T - W$ the budget left after the warm-up phase. We start by showing that Stratified SVARM is not afflicted with any bias.

Theorem 5. *The Shapley value estimate $\hat{\phi}_i$ of any $i \in \mathcal{N}$ obtained by Stratified SVARM is unbiased, i.e.,*

$$\mathbb{E}[\hat{\phi}_i] = \phi_i.$$

Next, we consider the variance of the Shapley value estimates and quickly introduce some notation. Let $A_{i,\ell} \subseteq \mathcal{N}_i$ be a random coalition of size ℓ distributed with $\mathbb{P}(A_{i,\ell} = S) = \binom{n-1}{\ell}^{-1}$. Define the strata variances

$$\sigma_{i,\ell}^{+\,2} := \mathbb{V}\left[\nu(A_{i,\ell} \cup \{i\})\right] \text{ and } \sigma_{i,\ell}^{-\,2} := \mathbb{V}\left[\nu(A_{i,\ell})\right]. \quad (15)$$

Theorem 6. *The variance of any player's Shapley value estimate $\hat{\phi}_i$ obtained by Stratified SVARM is bounded by*

$$\mathbb{V}[\hat{\phi}_i] \leq \frac{2 \log n}{n\bar{T}} \sum_{\ell=1}^{n-3} \sigma_{i,\ell}^{+\,2} + \sigma_{i,\ell+1}^{-\,2}.$$

Together with the unbiasedness shown in Theorem 5, the variance bound implies the following MSE bound.

Corollary 2. *The MSE of any player's Shapley value estimate $\hat{\phi}_i$ obtained by Stratified SVARM is bounded by*

$$\mathbb{E}[(\hat{\phi}_i - \phi_i)^2] \leq \frac{2 \log n}{n\bar{T}} \sum_{\ell=1}^{n-3} \sigma_{i,\ell}^{+\,2} + \sigma_{i,\ell+1}^{-\,2}.$$

With our choice of the sampling distribution $\tilde{P}$ we achieved an easily interpretable bound on the MSE in which each stratum variance is equally weighted. Assuming that each stratum variance is bounded by some constant independent of n, the MSE bound in Corollary 2 is in $\mathcal{O}(\frac{\log n}{T - n \log n})$. Note that, by assumption, T is growing log-linearly with n so that the denominator is essentially driven by the asymptotics of T. Again, compared to existing theoretical results, with linear dependence on n, the logarithmic dependence on n is a significant improvement. Still, it is worth emphasizing that the more homogeneous strata with lower variances constitute the core improvement of Stratified SVARM, which are not reflected within the O-notation. Our first probabilistic bound is obtained by Chebyshev's inequality and the bound from Theorem 6.

Theorem 7. *The probability that the Shapley value estimate $\hat{\phi}_i$ of any fixed player $i \in \mathcal{N}$ deviates from ϕ_i by a margin of any fixed $\varepsilon > 0$ or greater is bounded by*

$$\mathbb{P}(|\hat{\phi}_i - \phi_i| \geq \varepsilon) \leq \frac{2 \log n}{\varepsilon^2 n\bar{T}} \sum_{\ell=1}^{n-3} \sigma_{i,\ell}^{+\,2} + \sigma_{i,\ell+1}^{-\,2}.$$

Lastly, our second probabilistic bound derived via Hoeffding's inequality is tighter, but less trivial. It requires some further notation, namely the ranges of the strata values:

$$r_{i,\ell}^{+} := \max_{S \subseteq \mathcal{N}_i : |S| = \ell} \nu(S \cup \{i\}) - \min_{S \subseteq \mathcal{N}_i : |S| = \ell} \nu(S \cup \{i\}), \quad (16)$$

$$r_{i,\ell}^{-} := \max_{S \subseteq \mathcal{N}_i : |S| = \ell} \nu(S) - \min_{S \subseteq \mathcal{N}_i : |S| = \ell} \nu(S). \quad (17)$$

Theorem 8. *The probability that the Shapley value estimate $\hat{\phi}_i$ of any fixed player $i \in \mathcal{N}$ deviates from ϕ_i by a margin of any fixed $\varepsilon > 0$ or greater is bounded by* $\mathbb{P}(|\hat{\phi}_i - \phi_i| \geq \varepsilon)$

$$\leq 2(n-3) \left(e^{-\frac{\bar{T}}{8n^2 (\log n)^2}} + 2 \frac{e^{-\Psi \left\lfloor \frac{\bar{T}}{4n \log n} \right\rfloor}}{e^{\Psi} - 1} \right),$$

where $\Psi = 2\varepsilon^2 n^2 \Big/ \left(\sum_{\ell=1}^{n-3} r_{i,\ell}^{+} + r_{i,\ell+1}^{-} \right)^2$.

This bound is of order $\mathcal{O}(ne^{-\frac{T - n \log n}{n^2 (\log n)^2}})$ showing a slightly worse dependency on n compared to Theorem 4 due to the introduction of strata.

Empirical Results

To complement our theoretical findings, we evaluate our algorithms and its competitors on commonly considered synthetic cooperative games and explainable AI scenarios in which Shapley values need to be approximated. In particular, we select parameterless algorithms that do not rely on provided knowledge about the value function of the problem instance at hand, since ours do not either. Besides the sampling distribution $\tilde{P}$ over coalition sizes proposed for Stratified SVARM (S-SVARM), we also consider sampling with the simpler uniform distribution over all sizes from 2 to $n-2$ (S-SVARM uniform). In order to allow for a fair comparison with KernelSHAP, which samples coalitions without replacement, we include with S-SVARM$^+$ (uniform) an empirical version of S-SVARM without the warm-up that also samples without replacement to compensate for this underlying advantage (see Algorithm 8 in Appendix B), which obviously comes at the price of space complexity linear in T.

We run the algorithms multiple times on the selected game types and measure their performances by the mean squared error (MSE) averaged over all players and runs depending on a range of fixed budget values T. Measuring the approximation quality by the MSE requires the true Shapley values of the considered games to be available. These are either given by a polynomial closed-form solution for the synthetic games (see Section 6.1) or we compute them exhaustively for our explanation tasks (see Section 6.2). The results of our evaluation are shown in Figure 3 and are presented in more detail in Appendix F.

As already said, we judge the algorithms' approximation qualities in dependence on the spent budget (model evaluations) T instead of the consumed runtime. In fact, the algorithms differ in actual runtime. For example SVARM performs less arithmetic operations than Stratified SVARM since it does not update all players' estimates $\hat{\phi}_i^+$ or $\hat{\phi}_i^-$

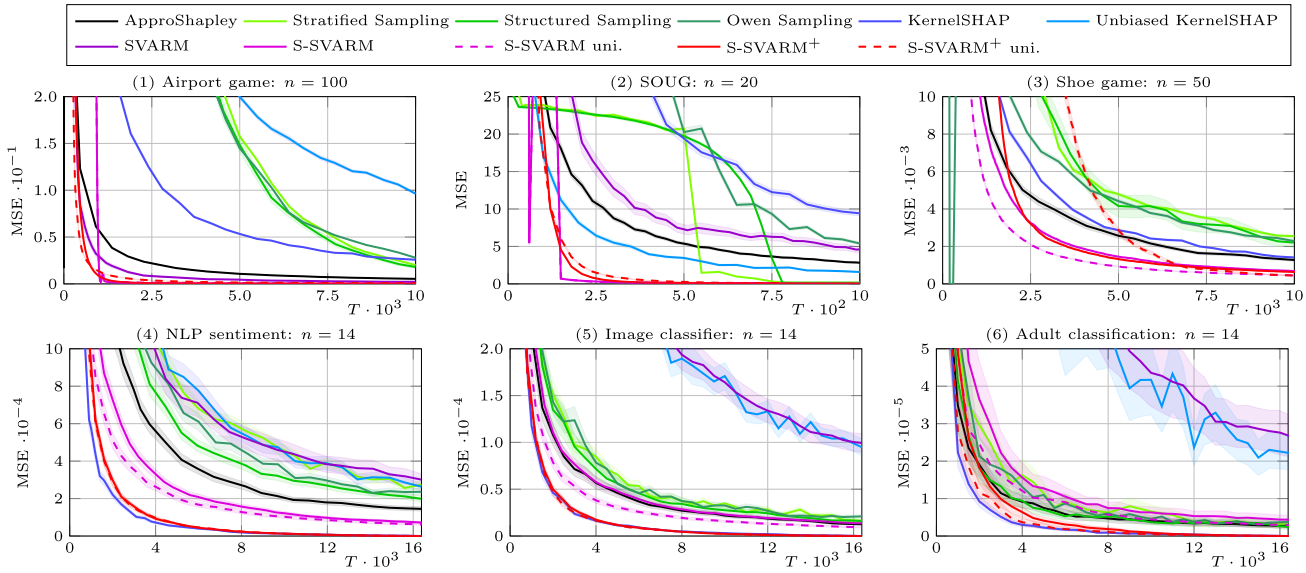


Figure 3: Averaged MSE and standard errors over 100 repetitions in dependence of fixed budget T: (1) Airport game, (2) Shoe game, (3) SOUG game, (4) NLP sentiment analysis, (5) Image classifier, (6) Adult classification.

with each sample. Some algorithms, e.g. KernelSHAP, vary strongly in their time consumption per sample since a costly quadratic optimization problem needs to be solved after observing all samples. We intentionally avoid the runtime comparison for three reasons: (i) the observed runtimes may differ depending on the actual implementation, (ii) the fixed-budget setting facilitates a coherent theoretical analysis where the observed information is restricted, (iii) evaluating the worth of a coalition poses the bottleneck in explanation tasks, rendering the difference in performed arithmetic operations negligible.

Synthetic Games

Cooperative games with polynomial closed-form solutions of their Shapley values are well suited for tracking the approximation error for large player numbers. We exploit this fact and investigate a broad range of player numbers n which are significantly higher than those for the explanation tasks. We conduct experiments on the predefined Shoe and Airport game as done in (Castro, Gómez, and Tejada 2009; Castro et al. 2017). Their degree of non-additivity poses a difficult challenge to all approximation algorithms. Further, we consider randomly generated Sum of Unanimity Games (SOUG) games (van Campen et al. 2018) which are capable of representing any cooperative game. The value function and Shapley values of each game are given in Appendix E.

We observe that S-SVARM itself already shows reliably good approximation performance across all considered games and budget ranges. It is significantly superior to its competitors ApproShapley and KernelSHAP and as expected, S-SVARM$^+$ extends the lead in approximation quality even more. In contrast, SVARM can rarely keep up with its refined counterpart S-SVARM. However, in light of the bounds on the MSEs in Corollary 1 and 2 this is not surprising: SVARM's MSE bound scales linearly with the variances ${\sigma_i^+}^2$ and ${\sigma_i^-}^2$ of *all* coalition values containing respectively not containing i, while the relevant variance terms ${\sigma_{i,\ell}^+}^2$ and ${\sigma_{i,\ell}^-}^2$ for S-SVARM are restricted to coalitions of fixed size. In most games, the latter terms are significantly lower since coalitions of the same size are plausibly closer in worth. Finally, S-SVARM is quite robust regarding the magnitude of the standard errors.

Explainability Games

We further conduct experiments on cooperative games stemming from real-world explainability scenarios, in particular, use cases in which local feature importance of machine learning models are to be quantified via Shapley values. The NLP sentiment analysis game is based on the DistilBERT (Sanh et al. 2019) model architecture and consists of randomly selected movie reviews from the IMDB dataset (Maas et al. 2011) containing 14 words. Missing features are masked in the tokenized representation and the value of a set is its sentiment score. In the image classifier game, we explain the output of a ResNet18 (He et al. 2016) trained on ImageNet (Deng et al. 2009). The images' pixels are summarized into $n = 14$ super-pixels and absent features are masked with mean imputation. The worth of a coalition is the returned class probability of the model (using only the present super-pixels) for the class of the original prediction which was made with all pixels being present. For the adult classification game, we train a gradient-boosted tree model on the adult dataset (Becker and Kohavi 1996). A coalition's worth is the predicted class probability of the true income class (income above or below $50\,000$) of the given datapoint with the absent features being removed via mean imputation. Since no polynomial closed-form solution exists for

the Shapley values in these games, we compute them exhaustively, limiting us to a feasible number of players for which we can track the MSE. While this restricts us to a player number (tokens, superpixels, features) of $n = 14$ due to limited computational resources, this is arguably still an appropriate and commonly appearing number of entities involved in an explanation task. We refer to Appendix E for a more detailed explanation of the chosen games.

A first observation is the close head-to-head race between S-SVARM$^+$ and KernelSHAP across the considered games leaving all other methods behind. Thus, S-SVARM$^+$ is the first sample-mean-based approach achieving rivaling state-of-the-art approximation quality. KernelSHAP's counterpart Unbiased KernelSHAP, designed to facilitate approximation guarantees similar to our theoretical results which KernelSHAP lacks, is clearly outperformed by S-SVARM. Given the consistency demonstrated by S-SVARM and S-SVARM$^+$, we claim that both constitute a reliable choice under absence of domain knowledge. We conjecture that the reason for the slight performance decrease of S-SVARM from synthetic to explainability games lies not only within the latent structure of ν, but is also caused by the lower player numbers. As our theoretical results indicate, its sample efficiency grows with n due to its enhanced update rule. However, conducting experiments with larger n becomes computationally prohibitive for explainability games, since the Shapley values have to be calculated exhaustively in order to track the approximation error. Further, our results indicate the robustness of S-SVARM$(^+)$ w.r.t. the utilized distribution $\tilde{P}$, which allows us to use the uniform distribution without performance loss, and secondly shows that our derived distribution is not just a theoretical artifact, but a valid contribution to express simpler bounds which are easier to grasp and interpret.

Conclusion

We considered the problem of precisely approximating the Shapley value of all players in a cooperative game under the restriction that the value function can be evaluated only a given number of times. We presented a reformulation of the Shapley value, detached from the ubiquitous notion of marginal contribution, facilitating the approximation by estimates of which a multitude can be updated with each access to the value function. On this basis, we proposed two approximation algorithms, SVARM and Stratified SVARM, which have a number of desirable properties. Both are parameter-free, incremental, domain-independent, unbiased, and do not require any prior knowledge of the value function. Further, Stratified SVARM shows a satisfying compromise between peak approximation quality and consistency across all considered games, paired with unmatched theoretical guarantees regarding its approximation quality. While fulfilling more desirable properties and not having to solve a quadratic optimization problem of size T in comparison to the state-of-the-art method KernelSHAP, effectively disabling on-the-fly approximations, our simpler sample-mean-based method Stratified SVARM$^+$ can fully keep up in common explainable AI scenarios, and even shows empirical superiority on synthetic games.

Limitations and Future Work. The quadratically growing number of strata w.r.t. n might pose a challenge for higher player numbers, which future work could remedy by applying a coarser stratification that assigns multiple coalition sizes to a single stratum. One could investigate the empirical behavior in further popular explanation domains such as data valuation, federated learning, or neuron importance and extend our evaluation to scenarios with higher player numbers. Since the true Shapley values are not accessible for larger n, a different measure of approximation quality than the MSE needs to be taken for reference. The convergence speed of the estimates is a naturally arising alternative. Our empirical results give further evidence for the non-existence of a universally best approximation algorithm and encourage future research into the cause of the observed differences in performance w.r.t. the game type. Further, it would be interesting to analyze whether structural properties of the value function, such as monotonicity or submodularity, have an impact on the approximation quality of both algorithms.

Acknowledgments

This research was supported by the research training group Dataninja (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia. We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824. We would like to thank Fabian Fumagalli and especially Patrick Becker for their efforts in supporting our implementation.

References

Ancona, M.; Öztireli, C.; and Gross, M. H. 2019. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 272–281.

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

Burgess, M. A.; and Chapman, A. C. 2021. Approximating the Shapley Value Using Stratified Empirical Bernstein Sampling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 73–81.

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.

Chen, H.; Covert, I. C.; Lundberg, S. M.; and Lee, S.-I. 2023. Algorithms to Estimate Shapley Value Feature Attributions. *Nature Machine Intelligence*, 5: 590–601.

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

Corder, K.; and Decker, K. 2019. Shapley Value Approximation with Divisive Clustering. In *18th IEEE International Conference On Machine Learning And Applications*, 234–239.

Covert, I.; and Lee, S.-I. 2021. Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In *24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, 3457–3465.

Covert, I.; Lundberg, S.; and Lee, S.-I. 2019. Shapley Feature Utility. In *Machine Learning in Computational Biology*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 248–255.

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

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

Ghorbani, A.; and Zou, J. Y. 2020. Neuron Shapley: Discovering the Responsible Neurons. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Illés, F.; and Kerényi, P. 2019. Estimation of the Shapley Value by Ergodic Sampling. *CoRR*, abs/1906.05224.

Lomeli, M.; Rowland, M.; Gretton, A.; and Ghahramani, Z. 2019. Antithetic and Monte Carlo Kernel Estimators for Partial Rankings. *Statistics and Computing*, 29(5): 1127–1147.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, 4768–4777.

Luo, C.; Zhou, X.; and Lev, B. 2022. Core, Shapley Value, Nucleolus and Nash Bargaining Solution: A Survey of Recent Developments and Applications in Operations Management. *Omega*, 110: 102638.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

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.

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.

O'Brien, G.; Gamal, A. E.; and Rajagopal, R. 2015. Shapley Value Estimation for Compensation of Participants in Demand Response Programs. *IEEE Transactions on Smart Grid*, 6(6): 2837–2844.

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

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

Rozemberczki, B.; and Sarkar, R. 2021. The Shapley Value of Classifiers in Ensemble Games. In *30th ACM International Conference on Information and Knowledge Management*, 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 *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 5572–5579.

Rubinstein, R. Y.; and Kroese, D. P. 2016. *Simulation and the Monte Carlo Method*. John Wiley & Sons.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *CoRR*, abs/1910.01108.

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

Simon, G.; and Vincent, T. 2020. A Projected Stochastic Gradient Algorithm for Estimating Shapley Value Applied in Attribute Importance. In *Machine Learning and Knowledge Extraction*, 97–115.

Soufiani, H. A.; Chickering, D. M.; Charles, D. X.; and Parkes, D. C. 2014. Approximating the Shapley Value via Multi-Issue Decompositions. In *Proceedings of the International conference on Autonomous Agents and Multi-Agent Systems*, volume 2.

Touati, S.; Radjef, M. S.; and Sais, L. 2021. A Bayesian Monte Carlo Method for Computing the Shapley Value: Application to Weighted Voting and Bin Packing Games. *Computers & Operations Research*, 125: 105094.

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.

Wang, J. T.; and Jia, R. 2023. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. In *26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *PMLR*, 6388–6421.

6

How Much Can Stratification Improve the Approximation of Shapley Values?

Author Contribution Statement

The author alone developed the idea, algorithms, analysis, experiment design, and visualization. Georg Haselbeck implemented and conducted experiments under the author's supervision. The author wrote the paper alone with the proofreading of both coauthors.

How Much Can Stratification Improve the Approximation of Shapley Values?

Patrick Kolpaczki[1( )], Georg Haselbeck[2], and Eyke Hüllermeier[2,3]

[1] Paderborn University, Paderborn, Germany
patrick.kolpaczki@upb.de
[2] University of Munich (LMU), Munich, Germany
[3] Munich Center for Machine Learning, Munich, Germany

Abstract. Over the last decade, the Shapley value has become one of the most widely applied tools to provide post-hoc explanations for black box models. However, its theoretically justified solution to the problem of dividing a collective benefit to the members of a group, such as features or data points, comes at a price. Without strong assumptions, the exponential number of member subsets excludes an exact calculation of the Shapley value. In search for a remedy, recent works have demonstrated the efficacy of approximations based on sampling with stratification, in which the sample space is partitioned into smaller subpopulations. The effectiveness of this technique mainly depends on the degree to which the allocation of available samples over the formed strata mirrors their unknown variances. To uncover the hypothetical potential of stratification, we investigate the gap in approximation quality caused by the lack of knowledge of the optimal allocation. Moreover, we combine recent advances to propose two state-of-the-art algorithms *Adaptive SVARM* and *Continuous Adaptive SVARM* that adjust the sample allocation on-the-fly. The potential of our approach is assessed in an empirical evaluation.

Keywords: Shapley Value · Cooperative Games · Explainable Artificial Intelligence · Feature Importance

1 Introduction

Over the last decade, machine learning models exhibited a significant increase in complexity, turning them eventually into non-transparent black boxes that seemingly resist any attempt to transfer their inner workings to a level of human comprehension. Meanwhile, developers are confronted with a recent rise in societal and legal pressure to ease understanding of their decision-making and thus provide trustworthiness, as for example the EU AI Act [1]. A common approach to deal with this rising demand is by providing post-hoc feature explanations. Additive feature explanations divide an observed numerical effect among the available features used by the applied model. This allows for interpreting the

L. Longo et al. (Eds.): xAI 2024, CCIS 2154, pp. 489–512, 2024.
https://doi.org/10.1007/978-3-031-63797-1_25

part assigned to a feature as its individual contribution to the effect of interest. Here, one commonly distinguishes between two feature explanation types. On the one hand, local explanations consider the model's prediction outcome for a single data point of interest as the effect to be split up, called feature attribution [20]. On the other hand, global explanations quantify the features' individual contributions to the generalization performance of the model on a chosen set of data points, also known as feature importance [8].

An established way to additively decompose an effect is by adopting a game-theoretical view. *Cooperative games* capture the spirit that features are agents or *players* which can form groups, called *coalitions*, and perform a task together, for which the group receives a numerical *reward*. Constructing this reward mechanism fittingly as the prediction value or the generalization performance elicits local, respectively global explanations. This in turn reduces the explanation task to finding an appropriate partition of the collective benefit obtained by all players cooperating together. The so far most popular solution to the problem of assigning fair payoffs is the Shapley value [29], since it is the only solution to provably fulfill desirable axioms that one would demand from such a partition. The Shapley value of a player can be understood as the weighted average of its *marginal contributions* to all coalitions, with the marginal contribution of a player to a coalition simply being the increase in received reward that the inclusion of that player causes. Unfortunately, it entails an inherent drawback. As its formula contains the rewards of all possible coalitions (feature subsets), the Shapley value's computational complexity is exponential w.r.t. the number of involved players in the game, turning quickly intractable in practice.

Current research on tackling this problem goes in two directions. By assuming a tightly restricted model type and reward mechanism, one stream of works reduces its computation to polynomial time for feature attribution [19,20]. However, these approaches are neither model-agnostic nor fruitful for feature importance. Even more importantly, the Shapley value found its way into numerous more areas, spanning from data valuation [12,16] to quantifying the contribution of base learners in ensembles [27] and neurons in deep networks [13], and well beyond machine learning such as for example economics [3]. For a broader overview in machine learning we refer to [28]. Hence it is of vital importance to tackle the problem on a more abstract and domain-independent level that allows to transfer solutions. The second stream does justice to this assessment and proposes to approximate the Shapley value with barely any underlying assumptions (see Sect. 2). Most approximation methods strike a balance between precision and approximation time by returning sample-based mean-estimates.

Among these, the technique of stratification has been employed by a number of algorithms. Stratification takes advantage of the observation that coalitions of the same size may tend to obtain similar rewards. Grouping them by size creates subpopulations, called *strata*, of higher homogeneity w.r.t. the coalitions' rewards than then population of the whole power set of players. The increase in homogeneity, or the reduction in variance in other words, speeds up sample-based mean estimation, as subestimates for each *stratum* converge faster. The key to

exploit this technique to its fullest is an allocation of available samples that prioritizes strata with higher variance. However, this comes with two hurdles. First, the optimal allocation has to be derived analytically after investigating the algorithm's precision depending on each stratum's variance and the number of samples spent on it. And second, the stratum variances are a priori unknown and can only be estimated. We call approaches oblivious to these variances with a fixed sample allocation *static*, and those that adjust their allocation during the approximation process by learning from the observed samples *adaptive*.

Contribution. Despite first refinements offered in the literature, the hypothetical potential of stratification using the optimal allocation is left unexplored. Assessing it would not only shed light on the gap that current stratifying methods have to close, but also reveal what performance improvements are to be expected at most. Moreover, adaptive methods have only been proposed for the class of approaches that sample marginal contributions, while the more recent class of sampling coalition-reward pairs *Stratified SVARM* [18] is being left untouched so far. Hence our contribution is divided into multiple parts:

- First, we reflect upon stratification for Shapley approximations and establish guiding terminology in Sects. 4 and 6.
- We derive the theoretically optimal sample allocation for the state-of-the-art *Stratified SVARM* algorithm in Sect. 5.
- By transferring and improving adaptive techniques, we propose the enhanced model-agnostic algorithms *(Continuous) Adaptive SVARM* in Sect. 7.
- Finally, we conduct an empirical evaluation of the benefit of adaptive compared to static stratification for both sampling approaches in Sect. 8.

2 Related Work

Interpreting the Shapley value as a weighted average of marginal contributions allows to also view it as an expectation of those and thus *ApproShapley* [6] approximates it by sampling marginal contributions, with further theoretical guarantees provided in [21]. Following, [21] introduced with *Stratified Sampling* the stratification by size, employing a static sample allocation over the strata. *Structured Sampling* [30] falls within this class too as it distributes the samples uniformly over all strata. The first algorithms to consider adaptive stratification were *Standard Deviation Sampling* [25] and *St-ApproShapley-opt* [5]. While the former, represents a multi-armed bandit-based philosophy, the latter chases the optimal allocation by estimating the strata's variances. A more sophisticated mechanism is employed by the *Stratified Empirical Bernstein Method* [4], which evaluates for each sample to be drawn next the most promising stratum. Further model-agnostic methods relying on the notion of marginal contributions are given in [15,22,26]. Shifting to sampling coalition-reward pairs, [7] divide the Shapley value into two sums, for which [18] propose and theoretically analyze *Stratified*

SVARM. More outstanding, yet popular is *KernelSHAP* [20]. It exploits the correspondence of the Shapley value to the solution of a weighted least square optimization problem, which it approximately constructs with each observed coalition, but does not allow to apply stratification.

3 The Shapley Value and Its Approximation

Cooperative games are formally represented by a tuple $(\mathcal{N}, \nu)$ comprising the set of players $\mathcal{N} = \{1, \ldots, n\}$, which we identify by natural numbers, and a value function $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ that assigns to each *coalition* $S \subseteq \mathcal{N}$ a real-valued *worth* $\nu(S)$. Despite its simplicity, this formalism possesses the capability of modelling countless scenarios in which human or possibly nonhuman entities can form groups to attain a collective benefit. Given the availability of the players in $\mathcal{N}$, the question of how to divide the worth $\nu(\mathcal{N})$ of the grand coalition $\mathcal{N}$ among all members in equitable manner arises. The Shapley value [29] provides a solution based on the notion of *marginal contributions*. We call the increase in worth $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$ caused by the inclusion of player i in presence of a coalition S its marginal contribution to S. The Shapley value of any player $i \in \mathcal{N}$ is given by a weighted average of its marginal contributions:

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

One can derive this formula by imposing the four axioms efficiency, symmetry, additivity, and the dummy-property (see [28] for further explanations), which capture a widely accepted intuition of fairness in the context of profit allocation. The Shapley value is provably the only solution to fulfill all of these axioms simultaneously [29]. This uniqueness arguably constitutes the key driver for its popularity in and outside of XAI. However, it comes with the major drawback of computational complexity. Without strong assumptions on the structure of ν, the exact computation of the Shapley value is NP-hard [10] because the number of subsets grows exponentially fast with the number of players n. Approximations are therefore needed to make the computation practically feasible.

We consider the *fixed-budget setting* in which all of the latent but unknown Shapley values $\phi_1, \ldots, \phi_n$ are to be approximated by estimates $\hat{\phi}_1, \ldots, \hat{\phi}_n$. The approximation algorithm is aware of $\mathcal{N}$, but has only restricted access to ν in the sense that the number of times it can evaluate the worth $\nu(S)$ of any S is limited by a fixed budget T. We judge the approximation quality by the *mean squared error* (MSE) averaged over all players:

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

The minimization of this measure is a widely demanded goal in the literature given its accessibility in theory as well as for empirical measurements.

4 Static Stratification

Since the Shapley value is a weighted mean of marginal contributions, it can also be seen as the expected value of a probability distribution over marginal contributions defined by the weights, thus, establishing mean estimates obtained via sampling as a natural approximation. Generally speaking, the precision of mean estimates, collected by randomly sampling from a distribution, depends on the distribution's variance. The lower the variance, the higher is the precision. Although the population to draw samples from is already fixed, *stratification* can still reduce the variance of the estimate. Instead of sampling from the whole population, one can form a partition, dividing the population into multiple subpopulations, called *strata*. By sampling from each *stratum* separately, mean estimates are obtained for all strata, and these are aggregated to the desired estimate of the whole population. This approach allows a more accurate approximation if the strata are reasonably homogeneous, or in other words, have significantly lower variances than the base population.

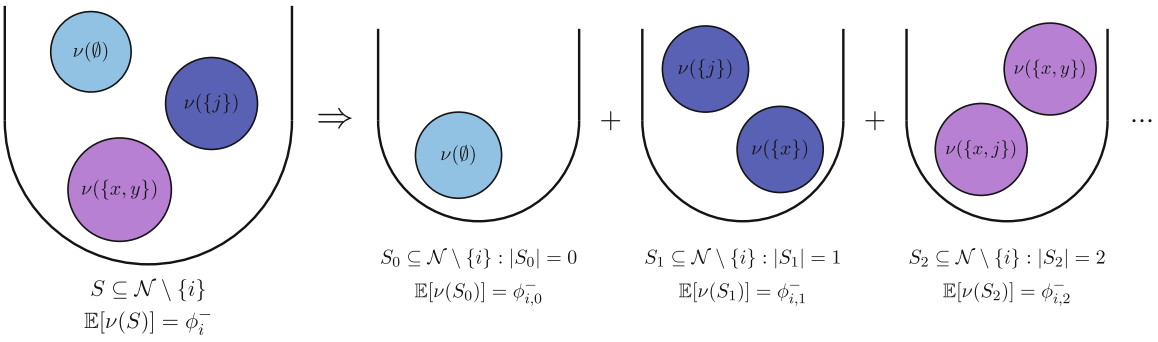


Fig. 1. Stratification at the example of Stratified SVARM: The population of coalitions $S \subseteq \mathcal{N} \setminus \{i\}$, depicted as marbles in an urn, is partitioned into multiple strata by grouping them by size. Aggregating the strata values $\phi_{i,\ell}^-$ yields the desired average ϕ_i^-. The homogeneity of each stratum increases the precision of sample-based estimates.

Stratified approaches have been proposed for two different representations of the Shapley value. First and most obviously, by viewing the individual marginal contributions of a player as elements of the population to be stratified, as first exemplified with *Stratified Sampling* by [21]. And second, [18] took advantage of a representation that dispenses with the ubiquitous notion of marginal contributions [2,7], based on averages of single coalition values. Stratifying this representation, [18] propose *Stratified SVARM*, a mechanism that updates the estimates of all players simultaneously with each sampled coalition.

We recapitulate both approaches in its simplest form proposed: *static stratification*. With this term, we coin the ignorance regarding the strata variances, which is reflected by an a priori allocation of samples over the strata. In other words, the mechanism to select the stratum to be sampled from next is oblivious

of the observations made and determined before the approximation task itself. We assume to sample with replacement as done by most methods for the sake of simplicity. All proofs of our theoretical results are given in Appendix B.

4.1 Stratified Sampling

Stratified Sampling proposed in [21] is the first algorithm to approximate a player's Shapley value based on stratification. The marginal contributions $\Delta_i(S)$ of a fixed player i are grouped by the cardinality of S into n many strata. Fittingly, the definition of the Shapley value uses this partition also from an algebraic view, suggesting a convenient interpretation. For any $\ell \in \mathcal{L}_\Delta := \{0, \ldots, n-1\}$, we define the ℓ-th stratum value of player i as the average of its marginal contributions to all coalitions of size ℓ, i.e.

$$\phi_{i,\ell} := \frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \{i\} \\ |S| = \ell}} \Delta_i(S)\,. \tag{3}$$

Next, by building upon this notion, we retrieve the Shapley value as the uniform average of the n many strata values:

$$\phi_i = \frac{1}{n} \sum_{\ell=0}^{n-1} \phi_{i,\ell}\,. \tag{4}$$

The sampling procedure maintains an estimate $\hat{\phi}_{i,\ell}$ for each stratum, given by the empirical mean of the independently sampled marginal contributions from the according stratum. Note that each sampled marginal contribution $\Delta_i(S)$ consumes two budget tokens as both values $\nu(S)$ and $\nu(S \cup \{i\})$ need to be evaluated. The stratum value estimates are aggregated in the same manner as the true values in order to obtain the final Shapley value estimate $\hat{\phi}_i$.

Since each player has its own population of marginal contributions, we are dealing with n separate approximation problems. In [21], the authors do not specify how to approximate the Shapley values of all players, though one can simply divide the available budget T among them and repeat the procedure for each player. We denote by $m_{i,\ell}$ the number of sampled marginal contributions from player i's ℓ-th stratum after the algorithm's termination and call $(m_{i,\ell})_{i \in \mathcal{N}, \ell \in \mathcal{L}_\Delta}$ the *sample allocation*, capturing the sample numbers of all n^2 many strata. Worth mentioning is the possibility that the sample numbers do not have to be fixed beforehand. Instead, Stratified Sampling employs a *static* sample allocation: each $m_{i,\ell}$ is fixed upon initialization. Given the variance $\sigma^2_{i,\ell} := \mathbb{V}[\Delta_i(S)]$ w.r.t. the uniform probability distribution over all coalitions $S \subseteq \mathcal{N} \setminus \{i\}$ of size ℓ we can express the achieved MSE.

Theorem 1. *The mean squared error of Stratified Sampling using any static sample allocation* $(m_{i,\ell})_{i \in \mathcal{N}, \ell \in \mathcal{L}_\Delta}$ *with* $m_{i,\ell} \geq 1$ *for all* $i \in \mathcal{N}$ *and* $\ell \in \mathcal{L}_\Delta$ *is*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{\ell=0}^{n-1} \frac{\sigma^2_{i,\ell}}{m_{i,\ell}}\,.$$

The condition $m_{i,\ell} \geq 1$ for all strata implies that each stratum gets assigned at least one sample which is necessary to avoid any bias of the final estimate $\hat{\phi}_i$, i.e. $\mathbb{E}[\hat{\phi}_i] = \phi_i$. In fact, under this condition even the strata estimates themselves are unbiased. For technical reasons, [21] specifies an allocation that increases with coalition size, irrelevant for our objective of MSE minimization. Instead, one could simply distribute the budget uniformly over all players and strata by assigning $m_{i,\ell} = \frac{T}{2n^2}$, leading to an MSE of

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] = \frac{2}{nT}\sum_{i=1}^{n}\sum_{\ell=0}^{n-1}\sigma_{i,\ell}^2\,. \tag{5}$$

For the sake of simplicity, we assume that the assigned fraction to $m_{i,\ell}$ is a natural number and save ourselves the effort of rounding the sample allocation otherwise. Even further, one can save some budget by taking into account that the strata $\ell = 0$ and $\ell = n-1$ contain exactly one marginal contribution, reducing the variances $\sigma_{i,0}^2$ and $\sigma_{i,n-1}^2$ to zero for all i. The shared coalition values $\nu(\emptyset)$ and $\nu(\mathcal{N})$ can be reused saving a budget of $2n-2$. Hence, one could set $m_{i,0} = m_{i,n-1}$ to 1 for all players and split the remaining budget evenly, i.e. $m_{i,\ell} = \frac{\tilde{T}}{2n(n-2)}$ for all $\ell \in \mathcal{L}'_\Delta := \{1, \ldots, n-2\}$ with $\tilde{T} := T - 2n - 2$.

4.2 Stratified SVARM

Departing from marginal contributions, the second representation suitable for stratification separates the Shapley value into two sums ϕ_i^+ and ϕ_i^-:

$$\phi_i = \underbrace{\sum_{S\subseteq\mathcal{N}\setminus\{i\}} \frac{1}{n\cdot\binom{n-1}{|S|}}\nu(S\cup\{i\})}_{=:\phi_i^+} - \underbrace{\sum_{S\subseteq\mathcal{N}\setminus\{i\}} \frac{1}{n\cdot\binom{n-1}{|S|}}\nu(S)}_{=:\phi_i^-}\,. \tag{6}$$

Instead of approximating each sum separately, both weighted averages of coalition values are further stratified by the size of S in [18], illustrated in Fig. 1. For each player i and size $\ell \in \mathcal{L}_\Delta$, the following strata values are thus obtained:

$$\phi_{i,\ell}^+ := \frac{1}{\binom{n-1}{\ell}}\sum_{\substack{S\subseteq\mathcal{N}_i\\|S|=\ell}}\nu(S\cup\{i\}) \quad \text{and} \quad \phi_{i,\ell}^- := \frac{1}{\binom{n-1}{\ell}}\sum_{\substack{S\subseteq\mathcal{N}_i\\|S|=\ell}}\nu(S)\,. \tag{7}$$

The interpretation is simple and appealing: Each stratum value $\phi_{i,\ell}^+$ is the uniform average of all coalitions of size $\ell+1$ that include i. An analogous statement can be made for $\phi_{i,\ell}^-$. The Shapley value can then be written as

$$\phi_i = \frac{1}{n}\sum_{i=0}^{n-1}\phi_{i,\ell}^+ - \frac{1}{n}\sum_{i=0}^{n-1}\phi_{i,\ell}^-\,. \tag{8}$$

Update Mechanism. In contrast to *Stratified Sampling*, which samples from each stratum separately, *Stratified SVARM* reuses each sampled coalition A to update at least one stratum estimate $\hat{\phi}^+_{i,\ell}$ or $\hat{\phi}^-_{i,\ell}$ of each player, thereby getting the most out of the information observed. This is made possible by the observation that for any player i and any player subset A, $\nu(A)$ is either a part of $\phi^+_{i,|A|-1}$ if $i \in A$, or otherwise a part of $\phi^-_{i,|A|}$ if $i \notin A$. As a consequence, this effectively reduces the number of strata to sample from to $n+1$, one for each subset size, captured by $\mathcal{L}_\nu := \{0, \ldots, n\}$. In each time step a subset size to sample from is chosen and the sampled coalition of that size is used for the update mechanism. Note that in comparison to the sampling of marginal contributions, here each sample only consumes one budget token instead of two.

Warmup. Similar to Sect. 4.1 we define $m^+_{i,\ell}$ and $m^-_{i,\ell}$ as the number of sampled coalitions used to update $\phi^+_{i,\ell}$, or $\hat{\phi}^-_{i,\ell}$ respectively, for each $i \in \mathcal{N}$ and $\ell \in \mathcal{L}_\nu$. As we proposed in Sect. 4.1 for Stratified Sampling, [18] made use of the fact that some strata contain only a few coalitions. Stratified SVARM computes the strata values $\phi^-_{i,0}, \phi^+_{i,0}, \phi^-_{i,1}, \phi^+_{i,n-2}, \phi^-_{i,n-1}, \phi^+_{i,n-1}$ exactly for all players by evaluating all coalitions of size 0, 1, $n-1$, and n at the price of $2n+2$ budget tokens. This reduces the number of subset sizes to choose from to $n-3$. To guarantee unbiasedness, [18] introduce a warmup procedure preceding the sampling such that each of the remaining $2n^2 - 2n$ strata is covered by at least one sample, i.e. $m^+_{i,\ell}, m^-_{i,\ell} \geq 1$. It consumes a budget of $W := 2\sum_{s=2}^{n-2} \lceil \frac{n}{s} \rceil$ and we denote the remaining budget left for sampling as $\bar{T} := W + 2n + 2$.

Further, we let m_ℓ be the number of sampled coalitions of size ℓ after the warmup for each $\ell \in \mathcal{L}'_\nu := \{2, \ldots, n-2\}$. Note that although the numbers $m^+_{i,\ell}$ and $m^-_{i,\ell}$ are determining the approximation quality, the algorithm is only fixing an allocation on the level of the subset sizes, i.e. $(m_\ell)_{\ell \in \mathcal{L}'_\nu}$. Here, again, given a static sample allocation and the variances $\sigma^2_{i,\ell,+} := \mathbb{V}[\nu(S \cup \{i\})]$ and $\sigma^2_{i,\ell,-} := \mathbb{V}[\nu(S)]$ w.r.t. the uniform probability distribution over all coalitions $S \subseteq \mathcal{N} \setminus \{i\}$ of size ℓ, we can express a bound on the MSE.

Theorem 2. *The mean squared error of Stratified SVARM using any static sample allocation* $(m_\ell)_{\ell \in \mathcal{L}'_\nu}$ *is bounded by*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \leq \frac{1}{n^2}\sum_{\ell=2}^{n-2} \frac{1}{m_\ell} \sum_{i=1}^{n} \frac{\sigma^2_{i,\ell-1,+}}{\ell} + \frac{\sigma^2_{i,\ell,-}}{n-\ell}\,.$$

A subtle difference is that [18] selects the coalition size randomly according to a fixed probability distribution. The expected sample allocation can be interpreted as the chosen static allocation since it is not influenced by the observations made during sampling. Their proposed distribution is sophisticated, prioritizing subset sizes close to zero and close to n at the cost of those further in the middle of the spectrum in order to achieve a bound that equally weighs all strata variances. If

we instead split the remaining budget equally such that $m_\ell = \frac{\bar{T}}{n-3}$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \leq \frac{n-3}{n^2\bar{T}}\sum_{i=1}^{n}\sum_{\ell=2}^{n-2}\frac{\sigma^2_{i,\ell-1,+}}{\ell} + \frac{\sigma^2_{i,\ell,-}}{n-\ell}\,. \tag{9}$$

We demonstrate the sharpness of these bounds in Sect. 8.

5 Theoretically Optimal Allocation

The results shown above in Theorems 1 and 2 give rise to the question of how small the MSE can possibly be, and motivate the search for the responsible *optimal* sample allocation. For both sampling approaches *Stratified Sampling* and *Stratified SVARM*, the impact of a stratum's variance on the MSE is directly linked to its sample number. Hence, it comes quite naturally to fine-tune the sample allocation by adjusting it to the underlying variances. Obviously, these variances are unknown such that the resulting MSE can only be achieved in theory. However, this investigation will yield important insights, as we uncover the theoretical limit on the MSE for each approach that no static sample allocation can improve upon, thus showcasing the theoretical potential of stratification.

5.1 Optimal Stratified Sampling

In the pursuit of the optimal allocation, we minimize the MSE given in Theorem 1, while constraining the total number of evaluations according to the given budget. This can be formulated as the following optimization problem:

$$\begin{aligned}(m^*_{i,\ell})_{i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta} = \underset{(m_{i,\ell})_{i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta}}{\arg\min} \quad & \frac{1}{n^3}\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}\frac{\sigma^2_{i,\ell}}{m_{i,\ell}} \\ \text{s.t.} \quad & 2\sum_{i=1}^{n}\sum_{\ell=1}^{n-2} m_{i,\ell} = \tilde{T} \\ & m_{i,\ell} \in \mathbb{N} \qquad\qquad \forall i \in \mathcal{N}, \ell \in \mathcal{L}'_\Delta\end{aligned}$$

In order to allow for a fair comparison with *Stratified SVARM*, we assume that the strata $\ell = 0$ and $\ell = n-1$ are already computed exactly (see Sect. 4.1). While the constraint on the sample numbers reflects the fact that no fractions of samples can be taken, it also impedes the attempt to derive an analytical solution of the problem. For this reason, we allow non-negative real-valued numbers, i.e. substituting it by $m_{i,\ell} \in \mathbb{R}_{>0}\ \forall i \in \mathcal{N}, \ell \in \mathcal{L}$, and call the resulting optimization problem the *relaxed sample allocation problem for marginal contributions*.

Theorem 3. *The solution to the relaxed sample allocation problem for marginal contributions is given by the allocation*

$$m^*_{i,\ell} = \frac{\sigma_{i,\ell}}{2\sum_{j=1}^{n}\sum_{k=1}^{n-2}\sigma_{j,k}} \cdot \tilde{T} \quad \text{for all } i \in \mathcal{N}, \ell \in \mathcal{L}'_\Delta\,,$$

which yields the following mean squared error for Stratified Sampling:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i-\phi_i\right)^2\right]=\frac{2}{n^3\tilde{T}}\left(\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}\sigma_{i,\ell}\right)^2 .$$

Theorem 3 reveals that the optimal allocation partitions the budget among the strata in proportion to their fraction of the total sum of stratum standard deviations. This result is also known as the *Neyman allocation* [24]. It was discovered previously by [5] to improve *Stratified Sampling*, though used incorrectly by replacing the standard deviations with the variances. Some sample numbers might fall below 1, although at least one sample is needed per stratum. We fix this by rounding the sample allocation appropriately without violating the constraint.

5.2 Optimal Stratified SVARM

We now transfer the optimization problem in search for the optimal allocation from *Stratified Sampling* to *Stratified SVARM*. The objective function to be minimized is our result on the MSE in Theorem 2. Although this is only an upper bound, it does not significantly harm the meaningfulness of the solution to be derived, since the inequality stems from a minor technical detail introduced for the sake of readability. As a constraint, we impose again that the sum of samples drawn has to equal the budget left after the warmup:

$$\begin{aligned}(m_\ell^*)_{\ell\in\mathcal{L}'_\nu}=\min_{(m_\ell)_{\ell\in\mathcal{L}'_\nu}}\quad & \frac{1}{n^2}\sum_{\ell=2}^{n-2}\frac{1}{m_\ell}\sum_{i=1}^{n}\frac{\sigma_{i,\ell-1,+}^2}{\ell}+\frac{\sigma_{i,\ell,-}^2}{n-\ell}\\ \text{s.t.}\quad & \sum_{\ell=2}^{n-2}m_\ell=\bar{T}\\ & m_\ell\in\mathbb{N}\qquad\qquad\forall\ell\in\mathcal{L}'_\nu\end{aligned}$$

Again, we relax this optimization problem by allowing the sample numbers to be real-valued but non-negative, i.e. $m_\ell\in\mathbb{R}_{\geq 0}\ \forall\ell\in\mathcal{L}'_\nu$. We name this relaxation the *relaxed sample allocation problem for coalitions.*

Theorem 4. *The solution to the relaxed sample allocation problem for coalitions is given by the allocation*

$$m_\ell^*=\frac{\sqrt{\sum_{i=1}^{n}\frac{\sigma_{i,\ell-1,+}^2}{\ell}+\frac{\sigma_{i,\ell,-}^2}{n-\ell}}}{\sum_{k=2}^{n-2}\sqrt{\sum_{i=1}^{n}\frac{\sigma_{i,k-1,+}^2}{k}+\frac{\sigma_{i,k,-}^2}{n-k}}}\cdot\bar{T}\qquad \textit{for all } i\in\mathcal{N},\ell\in\mathcal{L}'_\nu\,,$$

which yields the following mean squared error for Stratified SVARM:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i-\phi_i\right)^2\right]\leq\frac{1}{n^2\bar{T}}\left(\sum_{\ell=2}^{n-2}\sqrt{\sum_{i=1}^{n}\frac{\sigma_{i,\ell-1,+}^2}{\ell}+\frac{\sigma_{i,\ell,-}^2}{n-\ell}}\right)^2 .$$

Similarly to Theorem 3, the optimal allocation for *Stratified SVARM* assigns each coalition size ℓ a number of samples that depends on the proportion its associated stratum variances contribute to the MSE bound. In contrast to the Neyman allocation, the update mechanism dilutes the relationship between strata and the sample numbers on the level of coalition sizes.

We dispense with an analytical comparison of the MSE given in Theorems 3 and 4 because the different variance notions of marginal contributions $\sigma^2_{i,\ell}$ and coalition values $\sigma^2_{\ell,+/-}$ of a specific game decide which approach offers the better approximation potential. Instead, we compare both optimal allocations empirically in Sect. 8. For the same reason, we consider asymptotic notation as inappropriate since it conceals the effect of small stratum variances, thus doing injustice to the core idea of stratification. Worth mentioning is that if all standard deviations are equal, both optimal allocations lead to the same MSE as the uniform allocations proposed in Sect. 4. The degree to which the former improve upon the latter increases with the variability of the standard deviations.

6 Adaptive Stratification

It comes at no surprise that the MSE achieved by *Stratified Sampling* and *Stratified SVARM* employing their respective optimal allocation (see Sect. 5) is not applicable in practice, since the stratum variances are unknown during approximation. Fortunately, overcoming this lack of information by estimating the stratum variances with the means of the observed samples poses a promising remedy, as it enables the adaptation of the sampling allocation during the approximation process itself. We call this *adaptive stratification*. In contrast to *static stratification*, the mechanism to select the stratum to be sampled from next is now *informed* and utilizes this (possibly inexact) knowledge.

Pursuing this idea, all adaptive methods are confronted with the *exploration-exploitation dilemma*. While adjusting the sample allocation according to the obtained knowledge about the strata variances promises to achieve a more accurate approximation, excessively exploiting it can even be harmful. Since it is based on the variances' estimates, it might lead to a poorly performing allocation if these are not approximated precisely enough. Hence, it is of vital importance to explore, i.e., to collect samples from strata with apparently low variances to ensure a convergence to the optimal allocation. Otherwise, the algorithm might be trapped in sampling too often from falsely assessed strata of high variance, without having the chance to correct its estimated optimal allocation by sampling more often from other strata. Performing exploration and exploitation to the right degree is key to successful adaptive stratification. We shortly recapitulate on current adaptive methods for *Stratified Sampling* in the following.

6.1 The Two-Phase Approach

One way to tackle the exploration-exploitation dilemma is by dividing both into two separate phases. The *Two-Staged-St-ApproShapley-opt* algorithm [5]

follows this motive as it samples in the first phase uniformly from all strata for exploration, and switches to exploitation in the second phase. This is done by estimating the optimal allocation w.r.t. the available budget T on the basis of Theorem 3 in combination with the observed strata variances in the exploration phase. Next, the algorithm calculates for each stratum the number of samples to be drawn in the exploitation phase such that the sum of both phases matches the estimated optimal allocation. Note that this is not guaranteed to be feasible, since sample numbers may already exceed their counterparts within the optimal allocation during exploration. Shortening the exploration phase reduces this risk, but unfortunately, the quality of the strata estimates as well. This concern is not dealt with in [5], where both phases are simply set to consume half of the total budget. We observe that the algorithm wrongly deviates from the Neyman allocation by considering the stratas' variances instead of their standard deviations.

6.2 Bandit-Based Approach

Instead of demanding a strict separation, one can transition from exploration towards exploitation in a more seamless manner. This paradigm is employed for the approximation of the Shapley value of a single player by *Standard Deviation Sampling* [25]. Instead of determining the sample numbers upfront it employs a probability distribution over the strata to select the next draw. The closer this distribution is to the proportions of the Newman allocation, the more exploitation is performed on average. On the other side, full exploration can be modelled by choosing a uniform distribution. The mix of both is achieved via a convex combination with $\epsilon(t) \in [0, 1]$ being the weight for the uniform distribution at timestep $t \in \{1, \ldots, T\}$, and respectively $1 - \epsilon(t)$ for the estimated optimal allocation. The degree of freedom to design the rate of the transition lies with the function ϵ. In [25], a sigmoid function with two parameters is chosen for controlling the percentage of exploration and the transition speed. Obviously, these are hyperparameters to be specified by the practitioner relying on domain-knowledge. In light of our aim to remain model-agnostic, this poses a considerable vulnerability to the robustness of the method.

6.3 Empirical Bernstein Bound

Continuing in the spirit of simultaneously performing exploration and exploitation, the *Stratified Empirical Bernstein Method* [4] combines both in an even more interwoven way. It relies on the Stratified Empirical Bernstein Bound [4], which bounds the probability that an estimate $\hat{\phi}_i$ deviates from ϕ_i by more than some specific degree. The innovation of the algorithm is to greedily select in each timestep the stratum that promises the highest reduction on the deviation probability given by the bound. As the actual sample numbers are also part of the considered degree of the deviation, the compromise between exploration and exploitation is steered automatically by the bound itself without further algorithmic intervention. In contrast to *Standard Deviation Sampling*, it comes with

a different flaw that impedes any effective usage in the absence of rich domain-knowledge: The bound assumes the ranges of all strata to be given, a requirement that is hard to fulfill for cooperative games stemming from explanation tasks.

7 Adaptive Stratification of SVARM

The previously presented techniques are only introduced for the class of methods which sample marginal contributions. Hence, we close this gap in the current literature by applying the two-phased approach to *Stratified SVARM*, as we consider it to be the only adaptive technique without critical deficiencies. This transfer yields *Adaptive SVARM*, and *Continuous Adaptive SVARM* in combination with a conceptual improvement, which we propose as new model-agnostic approximation algorithms for the Shapley value.

7.1 Adaptive SVARM

We present our algorithm (see Algorithm 1) formally building upon *Stratified SVARM* (see Sect. 4.2) and our derived optimal allocation in Sect. 5.2. The pseudocode of the subprocedures is given in Appendix A.

Warmup. To begin with, all coalitions of the sizes $0, 1, n-1, n$ are evaluated to compute all strata values $\phi^-_{i,0}, \phi^-_{i,1}, \phi^-_{i,n-1}, \phi^+_{i,0}, \phi^+_{i,n-2}, \phi^+_{i,n-1}$ exactly, captured by `ExactCalculation` (see Algorithm 4). Hence, *Adaptive SVARM* also maintains stratum estimates $\hat{\phi}^+_{i,\ell}$ for each $i \in \mathcal{N}$ and $\ell \in \mathcal{L}'_\nu$. Next, we keep valid estimates $\hat{\sigma}^2_{i,\ell,+/-}$ of the variances by ensuring to have observed at least two samples from each stratum. This is achieved by calling each procedure `WarmUp`$^+$ and `WarmUp`$^-$ (cf. Algorithm 6 and 7 in [18]) twice. In total this procedure leaves $\bar{T} := T - 2n - 2 - 2W$ budget tokens for the remaining phases.

Exploration Phase. In each timestep the size $s \in \{2, \ldots, n-2\}$ of the next coalition A to sample is determined in Round-robin fashion to mimic the uniform allocation. Afterwards, $A \subseteq \mathcal{N}$ is drawn u.a.r. The `Update` procedure (see Algorithm 3) evaluates the worth of A once and incrementally updates the stratum estimates $\hat{\phi}^{+/-}_{i,\ell}$ using the sample counters $c^{+/-}_{i,\ell}$. The computation of the variance estimates and $\hat{\sigma}^2_{i,\ell,+/-}$ is prepared by maintaining the sum of observed coalition values $\Sigma^{+/-}_{i,\ell}$ and the sum of squared values $\Omega^{+/-}_{i,\ell}$. We specify the length of the exploration phase by a parameter $\lambda \in [0, 1]$, representing the percentage of time steps it consumes of the available budget $\bar{T}$ after the warmup. Subsequently, the phase stretches over $\lambda\bar{T}$ many time steps. Assuming the latter to be a multiple of the number of remaining sizes $n-3$, each size is selected $\frac{\lambda\bar{T}}{n-3}$ many times.

Exploitation Phase. After completing the exploration phase, *Adaptive SVARM* has $(1-\lambda)\bar{T}$ many samples left for the exploitation phase, in which the previously employed uniform allocation is extended to the estimated optimal allocation over the budget $\bar{T}$. The procedure `CalculateAllocation` (see Algorithm 2) computes

Algorithm 1 Adaptive SVARM

Input: $\mathcal{N}$, $T \in \mathbb{N}, \lambda \in [0,1]$

$\hat{\phi}^+_{i,\ell}, \hat{\phi}^-_{i,\ell}, \hat{\sigma}^2_{i,\ell,+}, \hat{\sigma}^2_{i,\ell,-} \leftarrow 0$ for all $i \in \mathcal{N}, \ell \in \mathcal{L}_\nu$
$c^+_{i,\ell}, c^-_{i,\ell}, \Sigma^+_{i,\ell}, \Sigma^-_{i,\ell}, \Omega^+_{i,\ell}, \Omega^-_{i,\ell} \leftarrow 0$ for all $i \in \mathcal{N}, \ell \in \mathcal{L}_\nu$
$c_\ell \leftarrow 0$ for all $\ell \in \mathcal{L}_\nu$
ExactCalculation
2×WarmUp$^+$; 2×WarmUp$^-$
$\bar{T} \leftarrow T - 2n - 2 - 2W$
for $t = 1, \ldots, \lambda\bar{T}$ **do**
  $s \leftarrow (t - 1 \mod n - 3) + 2$
  Draw A from $\{S \subseteq \mathcal{N} \mid |S| = s\}$ uniformly at random
  Update(A)
end for
$(\hat{m}^*_\ell)_{\ell \in \{2,\ldots,n-2\}} \leftarrow$ CalculateAllocation
$t \leftarrow 1$
for $t \leq 1, \ldots, (1-\lambda)\bar{T}$ **do**
  $s \leftarrow \arg\min_{\ell \in \mathcal{L}'_\nu} \frac{c_s}{\hat{m}^*_s}$
  Draw A from $\{S \subseteq \mathcal{N} \mid |S| = s\}$ uniformly at random
  Update(A)
end for
$\hat{\phi}_i \leftarrow \frac{1}{n} \sum_{\ell=0}^{n-1} \hat{\phi}^+_{i,\ell} - \hat{\phi}^-_{i,\ell}$ for all $i \in \mathcal{N}$

Output: $\hat{\phi}_1, \ldots, \hat{\phi}_n$

the variance estimates on the basis of $\Sigma^{+/-}_{i,\ell}$ and $\Omega^{+/-}_{i,\ell}$, and plugs them into Theorem 4 to obtain the estimated optimal allocation:

$$\hat{m}^*_\ell = \frac{\sqrt{\sum_{i=1}^{n} \frac{\hat{\sigma}^2_{i,\ell-1,+}}{\ell} + \frac{\hat{\sigma}^2_{i,\ell,-}}{n-\ell}}}{\sum_{k=2}^{n-2} \sqrt{\sum_{i=1}^{n} \frac{\hat{\sigma}^2_{i,k-1,+}}{k} + \frac{\hat{\sigma}^2_{i,k,-}}{n-k}}} \cdot \bar{T} \qquad \text{for all } i \in \mathcal{N}, \ell \in \mathcal{L}'_\nu\,, \tag{10}$$

where $\hat{m}^*_\ell$ denotes the number of coalitions of size ℓ to be drawn in total in both phases. Let m^{explore}_ℓ and m^{exploit}_ℓ be the number of coalitions of size ℓ drawn during exploration, and exploitation respectively. Then, ideally, the sample number during exploitation is given by $m^{\text{exploit}}_\ell := \hat{m}^*_\ell - m^{\text{explore}}_\ell$. This is not necessarily feasible since one can not exclude $\hat{m}^*_\ell < m^{\text{explore}}_\ell$. In this case, m^{exploit}_ℓ is set to zero and after this altering of the allocation one has to adjust the other samples whose sum exceeds now $\bar{T}$. We perform this by excluding this coalition size from Eq. (10) and compute the allocation again, e.g., the summation in the denominator is taken without the affected size ℓ, while the budget is set to $\bar{T} - m^{\text{explore}}_\ell$. The intuition is to neglect the size ℓ, as more samples have been collected than demanded by the optimal allocation, and solve the optimization problem in Sect. 5.2 again with only the total budget available for the remaining sizes. Since $\hat{m}^*_\ell < m^{\text{explore}}_\ell$ can occur in the newly assigned allocation, this

procedure has to be repeated eventually. However, it is guaranteed to terminate due to n being finite. During each of the remaining $(1-\lambda)\bar{T}$ timesteps, the algorithm proceeds similarly to the exploration phase. With the purpose of sampling evenly, the only exception is that the next size s is chosen as the one with the lowest number of current samples c_s relative to its assignment $\hat{m}_s^*$. Finally, the stratum estimates are aggregated to obtain an estimate $\hat{\phi}_i$ for each player.

7.2 Continuous Adaptive SVARM

We find that the two-phase approach as presented so far does not fully exploit the collected information to estimate the optimal allocation as precisely as possible. While it derives variance estimates during the exploration phase, the observations during the exploitation phase remain untouched. Since these posses further useful information, we propose to update the estimated optimal sample allocation continuously during exploitation. The resulting algorithm *Continuous Adaptive SVARM* poses a straightforward improvement by additionally calling `CalculateAllocation` after each update (line 15) in the exploitation phase.

8 Empirical Results

In order to complement our theoretical work, we empirically compare the approximation qualities depending on the spent budget, assess the algorithms' efficacy, answer whether adaptive stratification improves upon its static counterpart, and quantify the remaining gap of *Stratified SVARM* variants to the theoretical optimum derived in Sect. 5.2. In particular, we are interested in a comparison between *ApproShapley* as a baseline of sampling marginal contributions without stratification, the class of algorithms sampling marginal contributions, including *Stratified Sampling* and its extension *Two-Staged-St-ApproShapley-opt* [5] (Adaptive Sampling), and on the contrary *Stratified SVARM* in combination with our proposed improvements *Adaptive SVARM* and *Continuous Adaptive SVARM*. In addition, we include for each class the optimal algorithm, that a priori knows all stratum variances, given by *Optimal Sampling* and *Optimal SVARM*. We consider cooperative games of different origin and structure: feature explanations (local and global) for real-world data and synthetic games.

Table 1. Used datasets and models for feature explanation tasks.

Type	Dataset	Features	Task	Model
Global	Adult census [17]	14	Classification	Random forest
Global	Bank Marketing [23]	16	Classification	Random forest
Global Local	Bike sharing [11]	12	Regression	Random forest XGBoost
Global	German credit [14]	20	Classification	Random forest
Local	ImageNet [9]	14	Classification	ResNet18

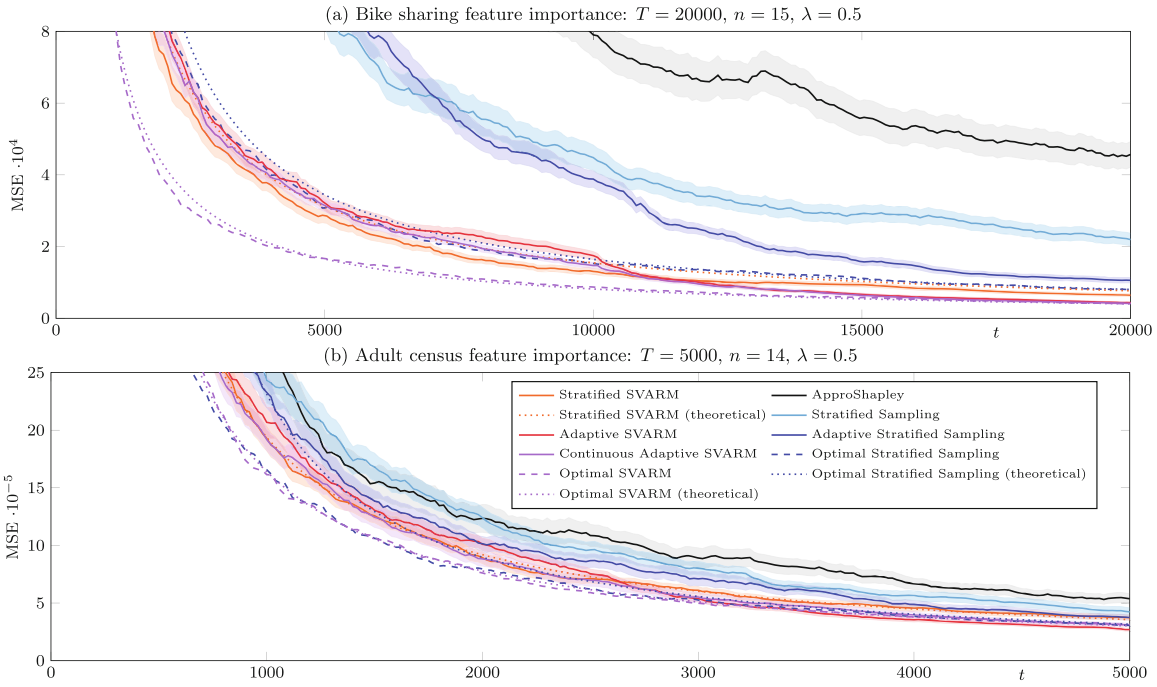


Fig. 2. Averaged MSE and standard error over 50 repetitions during approximation depending on current timestep t for global explanations. All adaptive algorithms use $\lambda = 0.5$. The performance of the hypothetical optimal algorithms (see Sect. 5) and theoretical bounds for Stratified SVARM (Eq. (9)), Optimal SVARM (Theorem 4), and Optimal Stratified Sampling (Theorem 3) are included.

Feature Importance Games. Given a model and dataset, we construct a cooperative game by setting the value function of a feature subset to be the classification accuracy (or R^2 for regression tasks) of the model on a test set (30% of datapoints). For each coalition S the model is fitted on the training set (70% of datapoints) with only the features contained in S. We use a default *sklearn* random forest consisting of 20 trees on all datasets given in Table 1.

Feature Attribution Games. For a model's prediction $h(x)$ on a specified datapoint x, the worth $\nu(S)$ of a feature subset is set to be the difference $h(x_S) - h(x_\emptyset)$, where $h(x_S)$ denotes the prediction with all features $\mathcal{N} \setminus S$ removed from x. For classification we take the class confidence $h_c(x_S) \in [0, 1]$ with $h(x) = c$ instead of the class label. Feature values in the Bike sharing dataset are removed by substituting them with their mean (or mode if categorical) within the dataset. For ImageNet pictures, semantic superpixels to be removed are replaced by their mean color. The used models and datasets are given in Table 1.

Synthetic Games. The experiments based on explanation tasks are limited to relatively low feature numbers because the Shapley values have to be computed in order to track the approximation error. Synthetic games [5,6,18] provide a possibility to investigate approximation algorithms for large player numbers n because their structure elicits a closed-form polynomial expression of the Shapley value.

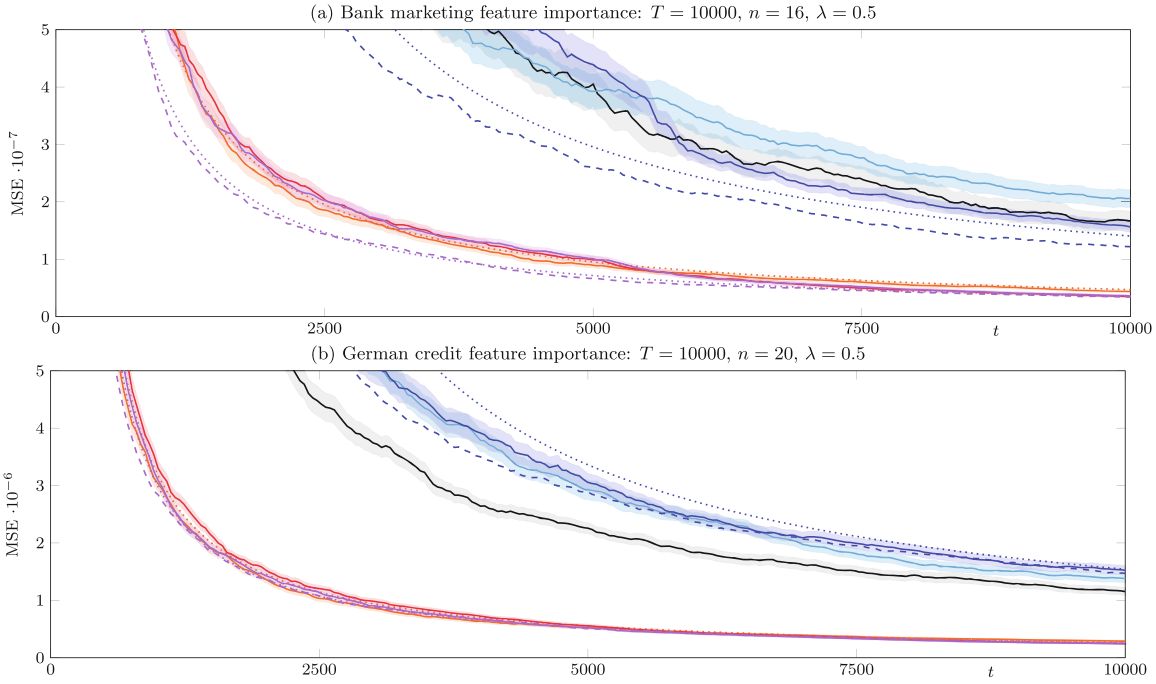


Fig. 3. Averaged MSE and std. error over 50 repetitions during approximation depending on current t for global explanations. For legend and further details see Fig. 2.

Our results show a significant gap between *Stratified Sampling* and its theoretical optimum across all datasets except for Fig. 3(b). Although not closing it completely, Adaptive SVARM reduces it visibly in Fig. 2(a) and Fig. 3(a) after switching to its exploitation phase. However, we also encounter cases in which it fails to have a positive effect on the approximation quality or performs even worse than the baseline *ApproShapley*. Interestingly, the deficit of the optimum to *ApproShapley* in Fig. 3(b) shows that stratifying marginal contributions can even be counterproductive to the goal of precise approximation.

On the other hand, the class of *Stratified SVARM* and our proposed improvements effortlessly outperform its counterpart based on marginal contributions except for feature attribution for regression tasks (see Fig. 4(a)). The optimality gap of *Stratified SVARM* is smaller or even nonexistent in Fig. 3(b). Independent of the gap's extent both of our algorithms close it after sufficient time spent in their exploitation phase. Hence, they not only improve upon state-of-the-art *Stratified SVARM*, but also lead us to conclude that we have reached optimality for this class of stratifying methods. To our surprise, we observe no considerable difference between *Adaptive SVARM* and *Continuous Adaptive SVARM* with $\lambda = 0.5$, implying that the variance estimates are precise enough and that λ can be further decreased. *Adaptive SVARM* reduces the approximation error at termination by (a) 28% and (b) 33% in Fig. 2, 18% in Fig. 4(b), and 70% in Fig. 5(b) compared to *Stratified SVARM*. Finally, the alignment of our theoretical statements with the empirical results, as especially exemplified for *Stratified SVARM* and Optimal SVARM, gives evidence for the precision of our analysis in Sect. 5 and the appropriateness of our approach.

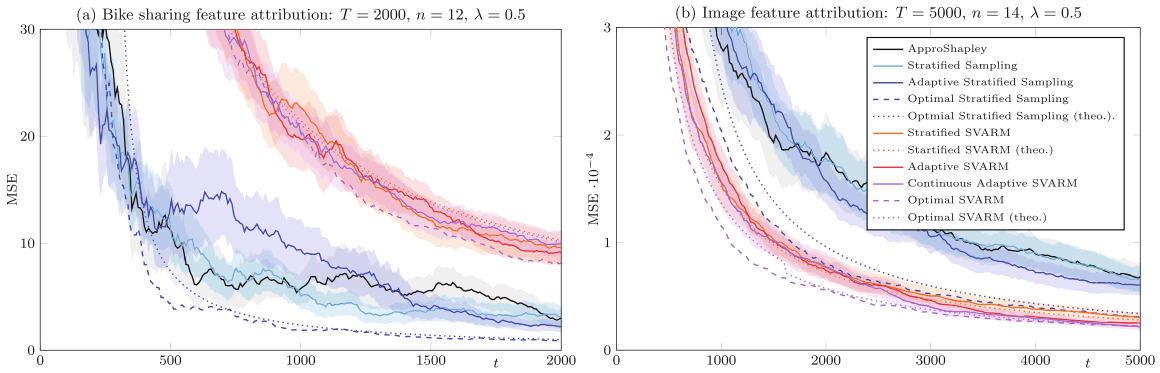


Fig. 4. Averaged MSE and standard error over 50 repetitions during approximation depending on current t for local explanations. All adaptive algorithms use $\lambda = 0.5$.

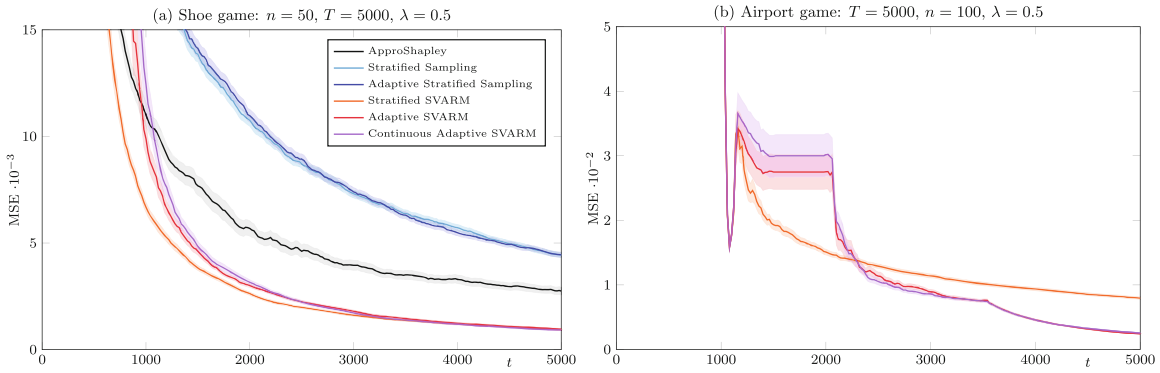


Fig. 5. Averaged MSE and standard error over 50 repetitions during approximation depending on current t for synthetic games. For further game details see [6,18].

9 Conclusion

We categorized stratified approaches for the approximation of all player's Shapley values in the fixed-budget setting by distinguishing between static and adaptive stratification. Recognizing the lack of more involved techniques for the class of methods that sample coalitions instead of marginals, we derived analytically the optimal allocation for *Stratified SVARM*. Moreover, we transferred the two-phase approach resulting in two novel approximation algorithms. Our empirical evaluation provides a multi-faceted insight. The gap of static methods to the theoretical optimum is of varies depending on the considered game. *Adaptive SVARM* and *Continuous Adaptive SVARM* close that gap and reach class optimality during approximation. The quadratically growing number of strata poses an inherent drawback of stratification for both classes. It prolongs the warmup, increases space complexity, and reduces the number of available samples per stratum. Future work could examine whether coarser strata encompassing multiple coalition sizes offer a reasonable workaround. Lastly, our work is effortlessly transferred to other semi-values proposed in game theory like the Banzhaf value.

Acknowledgments. This research was supported by the research training group Dataninja (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Pseudocode

CalculateAllocation begins by computing the variance estimates prepared by the observations during the exploration phase. Next, it estimates the optimal allocation in potentially multiple iterations. The sizes ℓ which are still part of the optimization problem are kept in $\mathcal{M}$. Each size whose sample number $\hat{m}_\ell^*$ does not exceed its number c_ℓ from the exploration phase, i.e. $\hat{m}_\ell^* \leq c_\ell$, is assigned after computation to the set of dropouts $\mathcal{F}$. The available budget $\bar{T}'$ considered is reduced by the sum of sample numbers of the current dropouts. We assume Round to return a vector of natural numbers by rounding appropriately.

Update updates the affected stratum estimates identical to [18] by iterating over all players. The variance estimation is prepared by maintaining for each stratum the sum observed coalition values in $\Sigma_{i,\ell}^{+/-}$ and the sum of squared values in $\Omega_{i,\ell}^{+/-}$. Note that the value function is only accessed once.

ExactCalculation iterates over all coalitions of size $0, 1, n-1, n$. Each coalition is used to call the Update procedure. After $2n+2$ spending budget tokens the strata values $\phi_{i,0}^-, \phi_{i,0^+}, \phi_{i,1}^-, \phi_{i,n-2}^+, \phi_{i,n-1}^-, \phi_{i,n-1}^+$ are computed exactly.

Algorithm 2 CalculateAllocation

$\hat{\sigma}^2_{i,\ell-1,+} \leftarrow \frac{1}{c^+_{i,\ell-1,}-1}\left(\Omega^+_{i,\ell-1} - \frac{1}{c^+_{i,\ell-1,}}\left(\Sigma^+_{i,\ell-1}\right)^2\right)$ for all $i \in \mathcal{N}, \ell \in \mathcal{L}'_\nu$

$\hat{\sigma}^2_{i,\ell,-} \leftarrow \frac{1}{c^-_{i,\ell,}-1}\left(\Omega^-_{i,\ell} - \frac{1}{c^-_{i,\ell,}}\left(\Sigma^-_{i,\ell}\right)^2\right)$ for all $i \in \mathcal{N}, \ell \in \mathcal{L}'_\nu$

$\hat{m}_\ell^* \leftarrow 0$ for all $\ell \in \{2, \ldots, n-2\}$

$\mathcal{M}, \mathcal{D} \leftarrow \{2, \ldots, n-2\}$

$\bar{T}' \leftarrow \bar{T}$

while $\mathcal{D} \neq \emptyset$ **do**

$$\hat{m}_\ell^* \leftarrow \frac{\sqrt{\sum_{i=1}^{n} \frac{\hat{\sigma}^2_{i,\ell-1,+}}{\ell} + \frac{\hat{\sigma}^2_{i,\ell,-}}{n-\ell}}}{\sum_{\ell \in \mathcal{M}} \sqrt{\sum_{i=1}^{n} \frac{\hat{\sigma}^2_{i,k-1,+}}{k} + \frac{\hat{\sigma}^2_{i,k,-}}{n-k}}} \cdot \bar{T}' \text{ for all } \ell \in \mathcal{M}$$

$\mathcal{D} \leftarrow \{\ell \in \mathcal{M} \mid \hat{m}_\ell^* \leq c_\ell\}$

$\bar{T}' \leftarrow \bar{T}' - \sum_{\ell \in \mathcal{D}} c_\ell$

$M \leftarrow M \setminus \mathcal{D}$

end while

$\hat{m}_\ell^* \leftarrow c_\ell$ for all $\ell \in \{0, \ldots, n-2\} \setminus \mathcal{M}$

Output: Round$(\hat{m}_0^*, \ldots, \hat{m}_{n-2}^*)$

Algorithm 3 UPDATE(A)

$v_A \leftarrow \nu(A)$; $c_{|A|} \leftarrow c_{|A|} + 1$
for $i \in A$ **do**
  $\hat{\phi}^+_{i,|A|-1} \leftarrow \frac{c^+_{i,|A|-1} \cdot \hat{\phi}^+_{i,|A|-1} + v_A}{c^+_{i,|A|-1}+1}$
  $c^+_{i,|A|-1} \leftarrow c^+_{i,|A|-1} + 1$
  $\Sigma^+_{i,|A|-1} \leftarrow \Sigma^+_{i,|A|-1} + v_A$
  $\Omega^+_{i,|A|-1} \leftarrow \Omega^+_{i,|A|-1} + (v_A)^2$
end for
for $i \in \mathcal{N} \setminus A$ **do**
  $\hat{\phi}^-_{i,|A|} \leftarrow \frac{c^-_{i,|A|} \cdot \hat{\phi}^-_{i,|A|} + v_A}{c^-_{i,|A|}+1}$
  $c^-_{i,|A|} \leftarrow c^-_{i,|A|} + 1$
  $\Sigma^-_{i,|A|} \leftarrow \Sigma^-_{i,|A|} + v_A$
  $\Omega^-_{i,|A|} \leftarrow \Omega^-_{i,|A|} + (v_A)^2$
end for

Algorithm 4 EXACTCALCULATION [18]

for $s \in \{0, 1, n-1, n\}$ **do**
  for $A \in \{S \subseteq \mathcal{N} \mid |S| = s\}$ **do**
    UPDATE(A)
  end for
end for

B Proofs

Static Allocation for Stratified Sampling. *Proof of Theorem 1*: Let $x^{(m)}_{i,\ell}$ be the m-th sampled marginal contribution of player i's ℓ-th stratum. Since $m_{i,\ell} \geq 1$ for all $i \in \mathcal{N}$ and $\ell \in \mathcal{L}_\Delta$, and the marginal contributions are drawn uniformly at random from their stratum, each estimate $\hat{\phi}_{i,\ell}$ and $\hat{\phi}_i$ is unbiased:

$$\mathbb{E}\left[\hat{\phi}_i\right] = \frac{1}{n}\sum_{\ell=0}^{n-1} \mathbb{E}\left[\hat{\phi}_{i,\ell}\right] = \frac{1}{n}\sum_{\ell=0}^{n-1} \mathbb{E}\left[\frac{1}{m_{i,\ell}}\sum_{m=1}^{m_{i,\ell}} x^{(m)}_{i,\ell}\right] = \frac{1}{n}\sum_{\ell=0}^{n-1} \phi_{i,\ell} = \phi_i\,.$$

Next, we investigate the variance where make use of the independent samples:

$$\mathbb{V}\left[\hat{\phi}_i\right] = \frac{1}{n^2}\sum_{\ell=0}^{n-1} \mathbb{V}\left[\hat{\phi}_{i,\ell}\right] = \frac{1}{n^2}\sum_{\ell=0}^{n-1} \frac{1}{m^2_{i,\ell}}\sum_{m=1}^{m_{i,\ell}} \mathbb{V}\left[x^{(m)}_{i,\ell}\right] = \frac{1}{n^2}\sum_{\ell=0}^{n-1} \frac{\sigma^2_{i,\ell}}{m_{i,\ell}}\,.$$

The bias-variance decomposition allows to combine both intermediate results and quantify the MSE for a single player. Averaging over all players yields:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}\left[\hat{\phi}_i\right] - \phi_i\right)^2 + \mathbb{V}\left[\hat{\phi}_i\right] = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{\ell=0}^{n-1} \frac{\sigma^2_{i,\ell}}{m_{i,\ell}}\,.$$

□

Static Allocation for Stratified SVARM. *Proof of Theorem 2*: The unbiasedness of each estimate $\hat{\phi}_{i,\ell}^{+}$, $\hat{\phi}_{i,\ell}^{-}$, and $\hat{\phi}_i$ has been shown in Lemma 8, Lemma 9, and Theorem 5 [18]. Let $\bar{m}_{i,\ell,+}, \bar{m}_{i,\ell,-}$ be the number of times the estimate $\hat{\phi}_{i,\ell}^{+}$ and respectively $\hat{\phi}_{i,\ell}^{-}$ got updated after the warmup and $m_{i,\ell,+}, m_{i,\ell,-}$ be the total numbers including the warmup.

Lemma 1. *For any $i \in \mathcal{N}$ and $\ell \in \mathcal{L}'_\nu$ the number of updates are binomially distributed with $\bar{m}_{i,\ell-1,+} \sim Bin\left(m_\ell, \frac{\ell}{n}\right)$ and $\bar{m}_{i,\ell,-} \sim Bin\left(m_\ell, \frac{n-\ell}{n}\right)$.*

The proof is analogous to Lemma 10 in [18]. We utilize Lemma 1 to bound the following expectations for all $i \in \mathcal{N}$ and all $\ell \in \mathcal{L}'_\nu$ (cf. Lemma 13 [18]):

$$\mathbb{E}\left[\frac{1}{m_{i,\ell-1,+}}\right] \leq \frac{n}{m_\ell \cdot \ell} \quad \text{and} \quad \mathbb{E}\left[\frac{1}{m_{i,\ell,-}}\right] \leq \frac{1}{\mathbb{E}\left[\bar{m}_{i,\ell,-}\right]} \leq \frac{n}{m_\ell(n-\ell)}\,.$$

From the proof of Theorem 6 in [18] we extract and reuse:

$$\begin{aligned}
\mathbb{V}\left[\hat{\phi}_i\right] &\leq \mathbb{E}\left[\frac{1}{n^2}\sum_{\ell=2}^{n-2}\frac{\sigma_{i,\ell-1,+}^2}{m_{i,\ell-1,+}} + \frac{\sigma_{i,\ell,-}^2}{m_{i,\ell,-}}\right] \\
&= \frac{1}{n^2}\sum_{\ell=2}^{n-2}\sigma_{i,\ell-1,+}^2 \cdot \mathbb{E}\left[\frac{1}{m_{i,\ell-1,+}}\right] + \sigma_{i,\ell,-}^2 \cdot \mathbb{E}\left[\frac{1}{m_{i,\ell,-}}\right] \\
&\leq \frac{1}{n}\sum_{\ell=2}^{n-2}\frac{1}{m_\ell}\left(\frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}\right)\,.
\end{aligned}$$

The bound on the variance enables us to take advantage of the bias-variance decomposition since the unbiasedness is already given:

$$\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] = \left(\mathbb{E}\left[\hat{\phi}_i\right] - \phi_i\right)^2 + \mathbb{V}\left[\hat{\phi}_i\right] \leq \frac{1}{n}\sum_{\ell=2}^{n-2}\frac{1}{m_\ell}\left(\frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}\right)\,.$$

Averaging the mean squared error over the players completes the proof:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \leq \frac{1}{n^2}\sum_{\ell=2}^{n-2}\frac{1}{m_\ell}\sum_{i=1}^{n}\frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}\,.$$

□

Optimal Allocation for Stratified Sampling. *Proof of Theorem 3*: The relaxed sample allocation problem for marginal contributions is given by:

$$\begin{aligned}
(m_{i,\ell}^*)_{i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta} = \underset{(m_{i,\ell})_{i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta}}{\arg\min} \quad & \frac{1}{n^3}\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}\frac{\sigma_{i,\ell}^2}{m_{i,\ell}} \\
\text{s.t.} \quad & 2\sum_{i=1}^{n}\sum_{\ell=1}^{n-2} m_{i,\ell} = \tilde{T} \\
& m_{i,\ell} \in \mathbb{R}_{>0} \qquad \forall i \in \mathcal{N}, \ell \in \mathcal{L}'_\Delta\,.
\end{aligned}$$

We derive the solution by employing Lagrangian multipliers, hence we minimize

$$L\left((m_{i,\ell})_{i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta},\lambda\right) = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}\frac{\sigma_{i,\ell}^2}{m_{i,\ell}} + \lambda\left(2\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}m_{i,\ell} - \tilde{T}\right).$$

We solve for $\nabla L = 0$ which requires

$$\frac{\partial}{\partial m_{i,\ell}}L = -\frac{\sigma_{i,\ell}^2}{n^3 m_{i,\ell}^2} + 2\lambda \overset{!}{=} 0 \;\forall i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta \quad \text{and} \quad \frac{\partial}{\partial\lambda}L = 2\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}m_{i,\ell} - \tilde{T} \overset{!}{=} 0\,.$$

This equation system yields together with the budget constraint

$$m_{i,\ell} = \frac{\sigma_{i,\ell}}{\sqrt{2\lambda n^3}} \quad \forall i\in\mathcal{N},\ell\in\mathcal{L}'_\Delta \quad \text{and} \quad \lambda = \frac{2}{n^3\tilde{T}^2}\left(\sum_{i=1}^{n}\sum_{\ell=1}^{n-2}\sigma_{i,\ell}\right)^2.$$

from which we derive the optimal allocation to minimize the objective function:

$$m_{i,\ell} = \frac{\sigma_{i,\ell}}{2\sum_{j=1}^{n}\sum_{k=1}^{n-2}\sigma_{j,k}}\cdot\tilde{T}\,.$$

□

Optimal Allocation for Stratified SVARM. *Proof of Theorem 4*: The relaxed sample allocation problem for coalitions is given by:

$$\begin{aligned}(m_\ell^*)_{\ell\in\mathcal{L}'} = \min_{(m_\ell)_{\ell\in\mathcal{L}'}} \quad & \frac{1}{n^2}\sum_{\ell=2}^{n-2}\frac{c_\ell}{m_\ell} \\ \text{s.t.} \quad & \sum_{\ell=2}^{n-2}m_\ell = \bar{T} \\ & m_\ell \in \mathbb{R}_{\geq 0} \qquad\qquad \forall\ell\in\mathcal{L}'_\nu\end{aligned}$$

where we define the coefficients $c_\ell := \sum_{i=1}^{n}\frac{\sigma_{i,\ell-1,+}^2}{\ell} + \frac{\sigma_{i,\ell,-}^2}{n-\ell}$. We derive the solution by employing Lagrangian multipliers, hence we minimize

$$L(m_2,\ldots,m_{n-2},\lambda) = \frac{1}{n^2}\sum_{\ell=2}^{n-2}\frac{c_\ell}{m_\ell} + \lambda\left(\sum_{\ell=2}^{n-2}m_\ell - \bar{T}\right).$$

We solve for $\nabla L = 0$ which requires

$$\frac{\partial}{\partial m_\ell}L = -\frac{c_\ell}{n^2 m_\ell^2} + \lambda \overset{!}{=} 0 \qquad \forall\ell\in\mathcal{L}'_\nu \quad \text{and} \quad \frac{\partial}{\partial\lambda}L = \sum_{\ell=2}^{n-2}m_\ell - \bar{T} \overset{!}{=} 0\,.$$

This equation system yields together with the budget constraint

$$m_\ell = \sqrt{\frac{c_\ell}{\lambda n^2}} \quad \forall\ell\in\mathcal{L}'_\nu \quad \text{and} \quad \lambda = \frac{1}{n^2\bar{T}^2}\left(\sum_{\ell=2}^{n-2}\sqrt{c_\ell}\right)^2.$$

from which we derive the optimal allocation to minimize the objective function:

$$m_\ell = \frac{\sqrt{c_\ell}}{\sum_{k=2}^{n-2} \sqrt{c_k}} \cdot \bar{T} = \frac{\sqrt{\sum_{i=1}^{n} \frac{\sigma^2_{i,\ell-1,+}}{\ell} + \frac{\sigma^2_{i,\ell,-}}{n-\ell}}}{\sum_{k=2}^{n-2} \sqrt{\sum_{i=1}^{n} \frac{\sigma^2_{i,k-1,+}}{k} + \frac{\sigma^2_{i,k,-}}{n-k}}} \cdot \bar{T}\,.$$

□

References

1. European commission. Ethics guidelines for trustworthy AI (2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
2. Ancona, M., Öztireli, C., Gross, M.H.: Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In: Proceedings of International Conference on Machine Learning (ICML), pp. 272–281 (2019)
3. Aumann, R.J.J.: Economic applications of the shapley value. In: Mertens, J.F., Sorin, S. (eds.) Game-Theoretic Methods in General Equilibrium Analysis. NATO ASI Serie, vol. 77, pp. 121–133. Springer, Dordrecht (1994). https://doi.org/10.1007/978-94-017-1656-7_12
4. Burgess, M.A., Chapman, A.C.: Approximating the shapley value using stratified empirical bernstein sampling. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp. 73–81 (2021)
5. Castro, J., Gómez, D., Molina, E., Tejada, J.: Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. Comput. Oper. Res. **82**, 180–188 (2017)
6. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. Comput. Oper. Res. **36**(5), 1726–1730 (2009)
7. Covert, I., Lundberg, S., Lee, S.I.: Shapley feature utility. In: Machine Learning in Computational Biology (2019)
8. Covert, I., Lundberg, S.M., Lee, S.: Understanding global feature contributions with additive importance measures. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
9. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
10. Deng, X., Papadimitriou, C.H.: On the complexity of cooperative solution concepts. Math. Oper. Res. **19**(2), 257–266 (1994)
11. Fanaee-T, H.: Bike Sharing. UCI Machine Learning Repository (2013) https://doi.org/10.24432/C5W894
12. Ghorbani, A., Zou, J.Y.: Data shapley: equitable valuation of data for machine learning. In: Proceedings of International Conference on Machine Learning (ICML), pp. 2242–2251 (2019)
13. Ghorbani, A., Zou, J.Y.: Neuron shapley: discovering the responsible neurons. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
14. Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994). https://doi.org/10.24432/C5NC77

15. Illés, F., Kerényi, P.: Estimation of the shapley value by ergodic sampling. CoRR abs/1906.05224 (2019)
16. Jia, R., et al.: Towards efficient data valuation based on the shapley value. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1167–1176 (2019)
17. Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), pp. 202–207 (1996)
18. Kolpaczki, P., Bengs, V., Muschalik, M., Hüllermeier, E.: Approximating the shapley value without marginal contributions. In: Proceeedings of AAAI Conference on Artificial Intelligence (AAAI), pp. 13246–13255 (2024)
19. Lundberg, S.M., Erion, G.G., Lee, S.: Consistent individualized feature attribution for tree ensembles. CoRR abs/1802.03888 (2018)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS), pp. 4768–4777 (2017)
21. Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., Rogers, A.: Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. CoRR abs/1306.4265 (2013)
22. Mitchell, R., Cooper, J., Frank, E., Holmes, G.: Sampling permutations for shapley value estimation. J. Mach. Learn. Res. **23**(43), 1–46 (2022)
23. Moro, S., Laureano, R., Cortez, P.: Using data mining for bank direct marketing: an application of the CRIDP-DM methodology. In: Proceedings of European Simulation and Modelling Conference (ESM) (2011)
24. Neyman, J.: On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J. Roy. Stat. Soc. **97**(4), 558–625 (1934)
25. O'Brien, G., Gamal, A.E., Rajagopal, R.: Shapley value estimation for compensation of participants in demand response programs. IEEE Trans. Smart Grid **6**(6), 2837–2844 (2015)
26. Okhrati, R., Lipani, A.: A multilinear sampling algorithm to estimate shapley values. In: Proceedings of International Conference on Pattern Recognition (ICPR), pp. 7992–7999 (2020)
27. Rozemberczki, B., Sarkar, R.: The shapley value of classifiers in ensemble games. In: Proceedings of ACM International Conference on Information and Knowledge Management (CIKM), pp. 1558–1567 (2021)
28. Rozemberczki, B., et al.: The shapley value in machine learning. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp. 5572–5579 (2022)
29. Shapley, L.S.: A value for n-person games. In: Contributions to the Theory of Games (AM-28), Volume II, pp. 307–318. Princeton University Press (1953)
30. van Campen, T., Hamers, H., Husslage, B., Lindelauf, R.: A new approximation method for the shapley value applied to the WTC 9/11 terrorist attack. Soc. Netw. Anal. Min. **8**(3), 1–12 (2018)

7

Comparing Shapley Value Approximation Methods for Unsupervised Feature Importance

Author Contribution Statement

The author made all contributions to this paper.

Proceedings of the DataNinja sAIOnARA 2024 Conference 13-15 DOI: 10.11576/dataninja-1158

Comparing Shapley Value Approximation Methods for Unsupervised Feature Importance

Patrick Kolpaczki PATRICK.KOLPACZKI@UPB.DE
Paderborn University, Germany

Abstract

Assigning importance scores to features is a common approach to gain insights about a prediction model's behavior or even the data itself. Beyond explainability, such scores can also be of utility to conduct feature selection and make unlabeled high-dimensional data manageable. One way to derive scores is by adopting a game-theoretical view in which features are understood as agents that can form groups and cooperate for which they obtain a reward. Splitting the reward among the features appropriately yields the desired scores. The Shapley value is the most popular reward sharing solution. However, its exponential complexity renders it inapplicable for high-dimensional data unless an efficient approximation is available. We empirically compare selected approximation algorithms for quantifying feature importance on unlabeled data.

Keywords: Shapley values, feature importance scores, unsupervised learning

1. Unsupervised Feature Importance

The increasing complexity of machine learning models as well as dimensionality of collected data is calling for a method to make both interpretable to the human user. A universally applicable approach are additive feature explanations which divide an observed numerical effect among the available features. Choosing this effect to be explained appropriately allows to interpret each feature's share as its contribution to the behavior of interest. In particular, the Shapley value [1] has emerged as the most frequently applied scoring rule. Popular examples include the features' contributions to a model's generalization performance [2, 3] and prediction value for a selected instance [4]. In the realm of unlabeled data and absence of a prediction model, Shapley-based feature importance scores have been utilized to perform dimensionality reduction [2]. Balestra et al. [5] refined this approach by proposing a feature ranking based on Shapley values that reduces redundancy among the selected features. Aiming at preserving the information contained in the data while minimizing correlation between the selected feature subset Balestra et al. employ the total correlation of shared by all all available features of the dataset as the numerical effect to be divided. For any subset S it is given by

$$C(S) = \sum_{X \in \mathcal{S}} H(X) - H(S) \tag{1}$$

where $H(X)$ and $H(S)$ denote the Shannon entropy of a single feature X and a set of features S respectively. This is made feasible by viewing the set of all feature values as observed realizations of a random variable.

2. Cooperative Games

A *cooperative game* is formally given by a pair $(\mathcal{N}, \nu)$ containing a finite set of *players* $\mathcal{N} = \{1, \dots, n\}$ and a *value function* $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ that assigns a real-valued *worth* to each *coalition* $S \subseteq \mathcal{N}$. This simple formalism is expressive enough to model feature subsets as coalitions that share some total correlation. The most popular solution to the question of how to divide the achieved worth $\nu(\mathcal{N})$ among all players is the *Shapley value* [1] as it is provably the only solution to fulfill certain axioms [1] that plausibly capture a notion of fairness. It assigns to each $i \in \mathcal{N}$ the share

$$\phi_i = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot [\nu(S \cup \{i\}) - \nu(S)] \tag{2}$$

and can be interpreted as a weighted average of marginal contributions $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$. Given the context of high-dimensional data yielding large player numbers, the computational complexity caused by the exponential number of coalitions renders any attempt to exactly calculate ϕ_i futile.

3. Shapley Value Approximation

The rapid increase of the Shapley value's popularity in recent years, spanning over various machine learning fields [6] and beyond, incentivized the research on how to approximate it, facilitating its practical usage. The approximation problem consists of the task of computing precise estimates $\hat{\phi}_1, \ldots, \hat{\phi}_n$ of all Shapley values with minimal resource consumption.

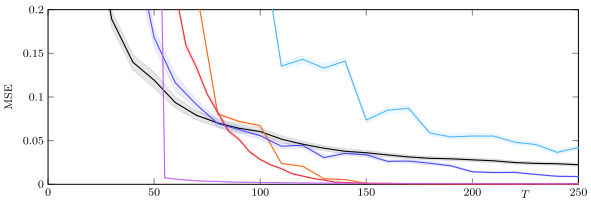
We consider the *fixed-budget setting* in which the number of times an approximation algorithm is allowed to access ν is limited by a *budget* $T \in \mathbb{N}$. This is motivated by the observation that the evaluation of large models or data poses a bottleneck, possibly even causing monetary costs when the access is provided remotely by another party. The quality of the estimates is measured by the mean squared error (MSE) averaged over all players which is to be minimized:

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

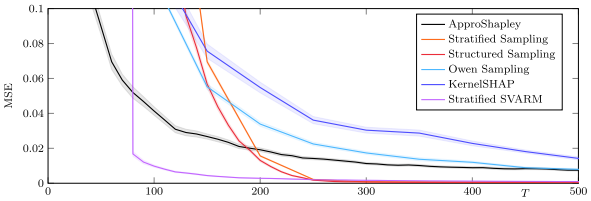
We shortly describe selected algorithms that we use for our experiments in Section 4. The first and simplest class of approximation methods leverages the fact that ϕ_i can be interpreted as player i's expected marginal contribution. This allows to obtain a mean estimate by randomly sampling marginal contributions. Castro et al. [7] propose with *ApproShapley* an algorithm that draws random permutations of $\mathcal{N}$. It extracts a marginal contribution of each player by iterating through a permutation. Following the spirit, *Stratified Sampling* [8] partitions the population of a player's marginal contributions into strata, each containing marginal contributions to coalitions S of the same size. This technique can increase estimation quality if $|S|$ has an influence on $\Delta_i(S)$. Closely related, *Structured Sampling* [9] modifies sampled permutations such that the marginal contributions to coalitions of different sizes appear in the same frequency. Departing from the discrete sum, *Owen Sampling* [10] updates an integral representation of the Shapley value [11]. Introducing another representation, Kolpaczki et al. [12] sample with *Stratified SVARM* single coalitions instead of marginal contributions. In combination with stratification it reaches higher sample efficiency as all players' estimates are updated with each coalition. Adopting a different view, *KernelSHAP* [4] solves a weighted least squares problem, filled by randomly drawn coalitions, of which the Shapley values are the solution.

4. Empirical Evaluation

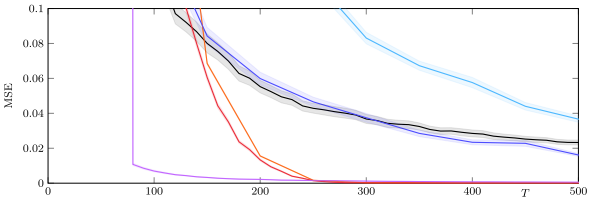
We compare the approximation quality of selected algorithms depending on the available budget T for unsupervised feature importance. In particular we use three real-world datasets: Breast Cancer, Big Five Personality Test, and FIFA 21 prepared as in [5]. A cooperative game is built from each dataset by interpreting the features as players and applying the total correlation as the corresponding coalition's worth. The approximation algorithms are run for a range of different budget values for multiple repetitions. In order to track the MSE, we calculate the Shapley values exhaustively beforehand. From Figure 1, *Stratified SVARM* emerges as significantly superior once it completes its warmup. *Stratified Sampling* and *Structured Sampling* perform on par or marginally better for higher budget ranges. The advantage of stratifying methods is likely to be caused by the impact of the feature subset size on the total correlation. In contrast, other methods including *KernelSHAP* perform clearly worse, except for *ApproShapley* displaying the lowest MSE given exteremely small budget.



(*a*) Breast Cancer: 9 features



(*b*) Big Five: 12 features



(*c*) FIFA 21: 12 features

Figure 1: Averaged MSE and std. error over 50 repetitions depending on available budget T.

Acknowledgments

This research was supported by the research training group Dataninja, funded by the German federal state of North Rhine-Westphalia.

References

[1] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, 1953.

[2] Shay B. Cohen, Eytan Ruppin, and Gideon Dror. Feature selection based on the shapley value. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 665–670, 2005.

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

[4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceeedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4768–4777, 2017.

[5] Chiara Balestra, Florian Huber, Andreas Mayr, and Emmanuel Müller. Unsupervised features ranking via coalitional game theory for categorical data. In *Proceedings of Big Data Analytics and Knowledge Discovery (DaWaK)*, pages 97–111, 2022.

[6] 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 International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5572–5579, 2022.

[7] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[8] 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.

[9] 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*, 8(3):1–12, 2018.

[10] Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate shapley values. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 7992–7999, 2020.

[11] Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):64–79, 1972. ISSN 00251909, 15265501.

[12] Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions. In *Proceeedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 13246–13255, 2024.

8 Shapley Value Approximation Based on k-Additive Games

Author Contribution Statement

Guilherme Pelegrina and the author jointly developed the idea of the paper while Guilherme Pelegrina majorly developed the algorithm. The author wrote the analysis by providing the proof and created the visualization. Both contributed equally to the experiment design. Guilherme Pelegrina implemented the algorithm and experiments, and the author implemented reference algorithms. Guilherme Pelegrina conducted all experiments. The paper was written by both while the author was majorly involved in editing and proofreading. Eyke Hüllermeier contributed by revising and proofreading.

Supplementary Material

An appendix to the paper is provided in Appendix B.

Shapley Value Approximation Based on k-Additive Games

Guilherme Dean Pelegrina [*1] **Patrick Kolpaczki** [*2 3] **Eyke Hüllermeier** [2 3]

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.

1. Introduction

The complexity of applied machine learning models experienced a rapid and certainly significant increase over the last decade. On the contrary, this development comes with an ever-rising burden to understand a model's decision-making, reaching a point at which the inner workings are 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 and developer. Among the 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 for a particular datapoint (Lundberg & Lee, 2017) or generalization performance on a test set (Covert et al., 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 et al., 2005; Marcílio & 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 (Peleg & Sudhölter, 2007), 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 the context of supply chain cooperation (Fiestras-Janeiro et al., 2011), the gain in reduction cost 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, measured by the Shapley value.

The applicability of the Shapley value exceeds by far the sphere of economics as its utility has been recognized by researchers of various disciplines. Most prominently, it has recently found its way into the branch of machine learning, especially as a model-agnostic approach, quantifying the importance of entities such as features, datapoints, and even model components like neurons in networks or base learners in ensembles (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 of interest. Shaping the measure of a coalition's worth adequately is pivotal to the informativeness of the importance scores obtained by the Shapley values. For ex-

[*]Equal contribution [1]Engineering School, Mackenzie Presbyterian University, São Paulo, Brazil [2]LMU Munich, Germany [3]MCML, Munich, Germany. Correspondence to: Guilherme Dean Pelegrina <guilherme.pelegrina@mackenzie.br>, Patrick Kolpaczki <patrick.kolpaczki@ifi.lmu.de>.

ample, 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 et al., 2020). Conversely, local feature attribution scores are obtained by splitting the model's prediction value for a fixed datapoint (Lundberg & Lee, 2017). The Shapley value's purpose is not limited to provide additive explanations since it has also been proposed to perform data valuation (Ghorbani & Zou, 2019), feature selection (Cohen et al., 2007), ensemble construction (Rozemberczki & Sarkar, 2021), and the pruning of neural networks (Ghorbani & Zou, 2020). Moreover, it has been applied to extract feature importance scores in several recent practical applications, such as in risk management (Nimmy et al., 2023), energy management (Cai et al., 2023), sensor array (re)design (Pelegrina et al., 2023b) and power distribution systems (Ebrahimi & Rastegar, 2024).

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 due to NP-hardness (Deng & Papadimitriou, 1994) 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 recently sharp increase in attention that XAI attracted, has rapidly fueled the research on approximation algorithms, leading to a diverse landscape of approaches (see (Chen et al., 2023)) for an overview related to feature attribution).

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:

- **(i)** *SVA*k_{ADD} fits a k-additive surrogate game to sampled coalitions, trying to mimick the given game by a simpler structure with a parameterizable degree of freedom while maintaining low representation error. The surrogate game's own Shapley values are obtained immediately due to its structure and yield precise estimates for the given game if the representation exhibits a good fit.
- **(ii)** *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.
- **(iii)** We prove the theoretical soundness of *SVA*k_{ADD} by showing analytically that its underlying optimization problem yields the Shapley value.
- **(iv)** 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.

The remainder of this paper is organized as follows. We describe existing works related to this paper in Section 2. Section 3 introduces the theoretical background behind our proposal. In Section 4, we present our novel approximation method. We conduct experiments for several real-world datasets in Section 5. Finally, in Section 6, we conclude our findings and highlight directions for future works.

2. Related Work

The problem of approximating the Shapley value, and the recent interest it attracted from various communities, lead to a multitude of diverse approaches to overcome its complexity. First to mention among the class of methods that can handle arbitrary games, without further assumptions on the structure of the value function, are those which construct mean estimates via random sampling. Fittingly, the Shapley value of each player can be interpreted as the expected marginal contribution to a specific probability distribution over coalitions. Castro et al. (2009) propose with *ApproShapley* the sampling of permutations from which marginal contributions are extracted. Further works, following this paradigm, employ stratification by coalition size (Maleki et al., 2013; Castro et al., 2017; van Campen et al., 2018; Okhrati & Lipani, 2020), 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 and updates the corresponding estimates with each sampled coalition, being further refined by *Adaptive SVARM* (Kolpaczki et al., 2024b). Guided by a different representation of the Shapley value, *KernelSHAP* (Lundberg & 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 & Lee, 2021) to be equivalent to importance sampling of single coalitions. Joining this family, (Pelegrina et al., 2023a) propose k_{ADD}-SHAP, which consists in a local explainability strategy that formulates the surrogate model assuming a k-additive game[1]. The authors locally adopt the Choquet

[1]Note that k_{ADD}-SHAP is limited to local explanations. In contrast, our proposed method *SVA*k_{ADD} differs by its applicability

integral as the interpretable model, whose parameters have a straightforward connection with the Shapley value.

On the contrary, tailoring the approximation to a specific application of interest 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 has proven to be a fruitful, resulting in multiple tailored approximation methods (Ghorbani & 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 involved players. Examples include the voting game (Bilbao et al., 2000) and the minimum cost spanning tree games (Granot et al., 2002) being used in operations research.

Besides the Shapley value's prominence for explaining the decision-making of a model, it has also found its way to more applied tasks. For instance, Nimmy et al. (2023) use it to quantify each feature's impact in predicting the risk degree in managing industrial machine maintenance, Pelegrina et al. (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 explanations. Moreover, Shapley values can be useful in feature engineering to perform feature selection. For instance, features with low relevance towards the model performance may be removed from the dataset without an impact onto the prediction quality (Pelegrina & Siraj, 2024).

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.

3.1. Cooperative Games and the Shapley Value

A cooperative game is formally described by n players, captured by the set $N = \{1, \dots, n\}$, 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

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

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 $\nu(\emptyset)$. Mathematically, efficiency means

$$\sum_{i=1}^{n} \phi_i = \nu(N) - \nu(\emptyset)\,. \tag{2}$$

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 & 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 re-trained 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 $\phi = (\phi_1, \dots, \phi_n)$ 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

to any formulation of a cooperative game. Moreover, in the context of explainable AI, it is capable of providing global explanations.

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:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right], \tag{3}$$

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 & Soneda, 1993; Grabisch, 1997a) by generalizing from marginal contributions to discrete derivatives. For any $S \subseteq N$ its Shapley interaction $I(S)$ is given by

$$I(S) = \sum_{A \subseteq N \setminus S} w_S \left(\sum_{A' \subseteq S} (-1)^{|S|-|A'|} \nu(A \cup A') \right) \tag{4}$$

with weights $w_S = \frac{(n-|A|-|S|)!|A|!}{(n-|S|+1)!}$. 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:

- If $I_{i,j} < 0$, there is a negative interaction (redundant effect) between features i and j.
- If $I_{i,j} > 0$, there is a positive interaction (complementary effect) between i and j.
- If $I_{i,j} = 0$, there is no interaction between i and 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, 1997a). Indeed, from the interactions one may easily retrieve the game payoffs by the following expression:

$$\nu(A) = \sum_{B \subseteq N} \gamma_{|A \cap B|}^{|B|} I(B)\,, \tag{5}$$

where $\gamma_{|A\cap B|}^{|B|}$ is defined by

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

are the Bernoulli numbers starting with $\eta_0 = 1$.

This linear transformation recovers any coalition value $\nu(A)$ by using the Shapley interaction values of all 2^n coalitions, thus including the Shapley values. Therefore, 2^n many 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 neglection of higher order interactions, by setting them to zero per default, comes naturally. For instance, Bordt & 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 fitted 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.

4.1. The k-Additive Optimization Problem

We leverage the representation of ν_k by means of interactions as given in Equation (5). 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 drop interactions of higher order than k:

$$\nu_k(A) = \sum_{\substack{B \subseteq N \\ |B| \leq k}} \gamma_{|A\cap B|}^{|B|} I^k(B)\,. \tag{6}$$

Note that, given 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

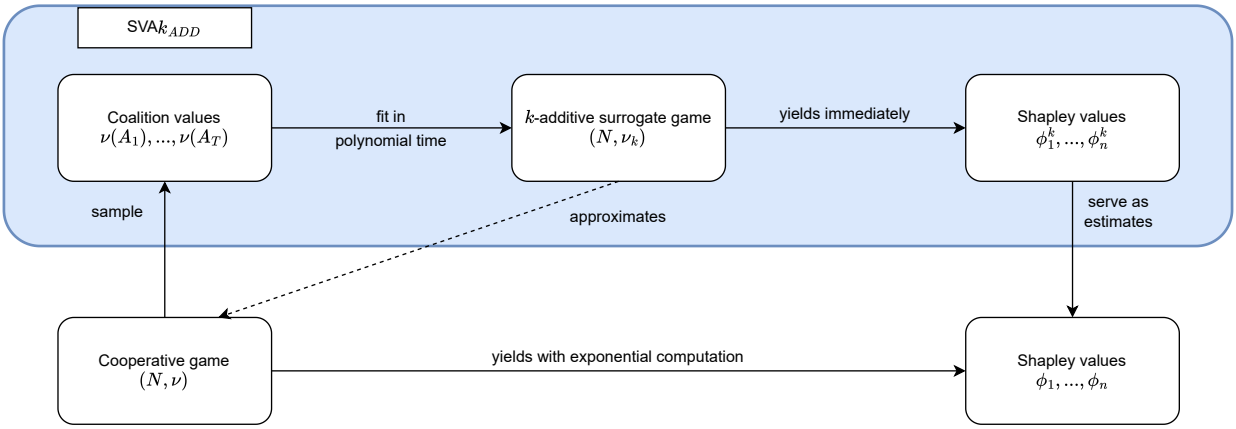


Figure 1: The sampled coalition values $\nu(A_1), \ldots, \nu(A_T)$ from the given game (N, ν) are used to fit a k-additive surrogate game (N, ν_k) in polynomial time. The Shapley values $\phi_1^k, \ldots, \phi_n^k$ of (N, ν_k) are obtained immediately from its k-additive representation. Since ν_k approximates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

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; 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 close as possible to ν and we therefore minimize the following expression, capturing by how much ν_k deviates from ν:

$$\sum_{A \in \mathcal{P}(N) \setminus \{\emptyset, N\}} w_A \left(\nu(A) - \nu_k(A)\right)^2 , \tag{7}$$

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

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

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 these weights coincide with those derived by Charnes et al. (1988) used in (Lundberg & Lee, 2017) for a different optimization problem. The result implies that having observed the cooperative game (N, ν) in its entirety with all coalitions contained, 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 its Shapley value. 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 challenge it poses. Worth mentioning is that the hardness of incorporating Shapley interactions of higher degree into weighted least squares optimizations has already 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 & Lee, 2017; Pelegrina et al., 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:

$$\begin{aligned} \min_{I^k} \quad & \sum_{A \in \mathcal{M} \setminus \{\emptyset, N\}} w_A \left(\nu(A) - \sum_{\substack{B \subseteq N \\ |B| \leq k}} \gamma_{|A \cap B|}^{|B|} I^k(B) \right)^2 \\ \text{s.t.} \quad & \nu(N) - \nu(\emptyset) = \sum_{\substack{B \subseteq N \\ |B| \leq k}} \left(\gamma_{|B|}^{|B|} - \gamma_0^{|B|} \right) I^k(B) \end{aligned} \tag{8}$$

To ensure the efficiency constraint, we force the sampling of $\emptyset$ and N. Each coalition $A \in \mathcal{P}(N) \setminus \{\emptyset, N\}$ is drawn according to an initial probability distribution p defined by

$$p_A = \frac{w_A^*}{\sum_{B \in N \setminus \{\emptyset, N\}} w_B^*} \,. \tag{9}$$

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 *SVAk*$_{\text{ADD}}$. 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 $\mathcal{P}(N)$, evaluates each, and extends $\mathcal{M}$ as well as $\nu_{\mathcal{M}}$. Finally, it solves the optimization problem in Equation (8) with weights w_A^* given by Theorem 4.2 (see Appendix B 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 the obtained solution during sampling when not all coalitions are observed. To the best of our knowledge, and it is also well-known, 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 & Lee, 2021).

Algorithm 1 *SVAk*$_{\text{ADD}}$

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: $I_1^k, \ldots, I_n^k$

5. Empirical Evaluation

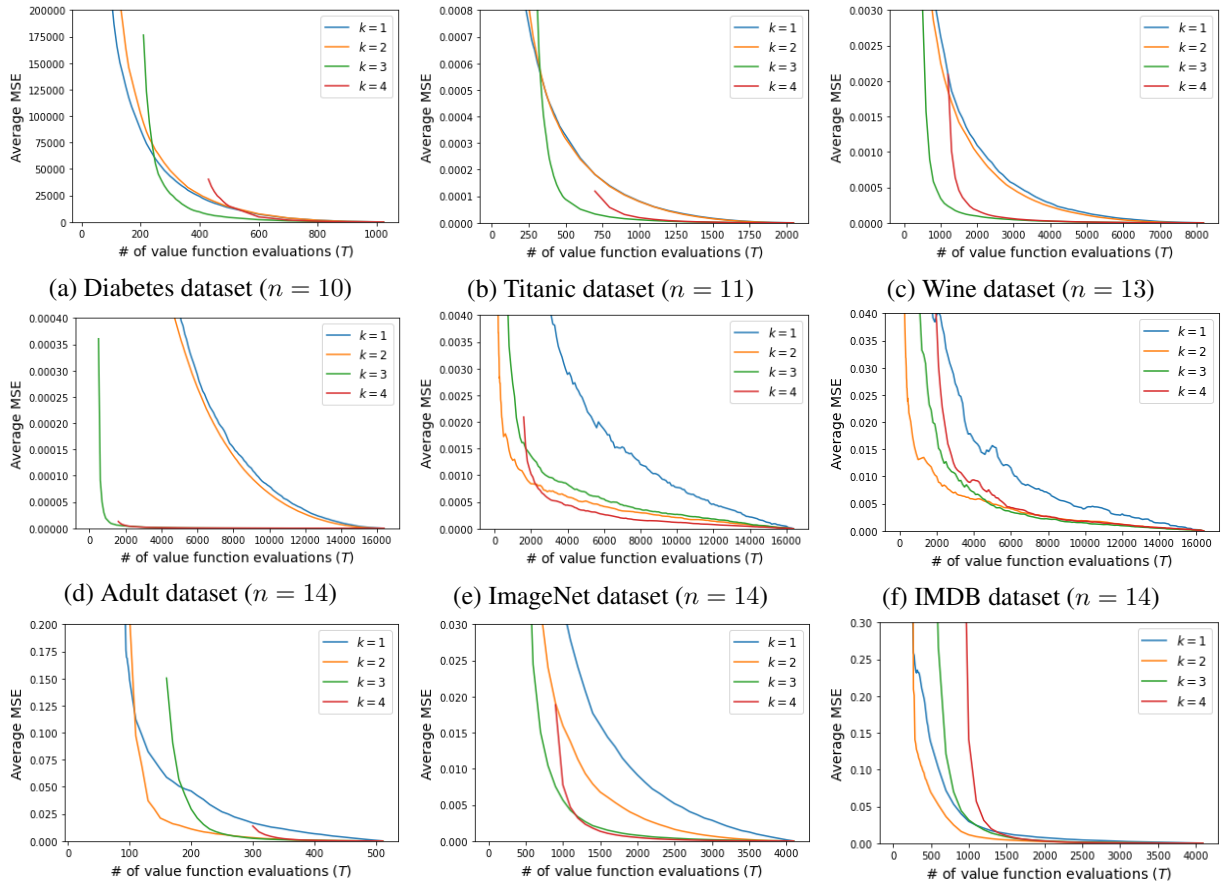
In order to assess the approximation performance of *SVAk*$_{\text{ADD}}$, we conduct experiments with cooperative games stemming from various explanation types. Although our method is not limited to a certain domain, we find the field of explainability best to illustrate its effectiveness. We consider several real datasets as well as different tasks. The evaluation of our proposal is mainly two-fold. Not only are we interested in the comparison of *SVAk*$_{\text{ADD}}$ 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 number of value function evaluations T, the obtained estimates $\hat{\phi}$ are compared with the Shapley values ϕ which we calculate exhaustively in advance. We measure approximation quality of the estimates by the mean squared error (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.

5.1. Datasets

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

Within global feature importance (Covert et al., 2020) the features' contributions to a model's generalization performance are quantified. This is done by means of accuracy



(a) Diabetes dataset ($n = 10$) (b) Titanic dataset ($n = 11$) (c) Wine dataset ($n = 13$)

(d) Adult dataset ($n = 14$) (e) ImageNet dataset ($n = 14$) (f) IMDB dataset ($n = 14$)

(g) Breast Cancer dataset ($n = 9$) (h) Big Five dataset ($n = 12$) (i) FIFA dataset ($n = 12$)

Figure 2: MSE of $SVAk_{\text{ADD}}$ averaged over 100 repetitions in dependence of available budget T for different additivity degrees k. Datasets stem from various explanation types: global (a)-(c), local (d)-(f), and unsupervised (g)-(i) with differing player numbers n.

for classification and the mean squared error for regression on a test set. For each evaluated coalition a random forest is retrained on a training set. We employ the *Diabetes* (regression, 10 features), *Titanic* (classification, 11 features), and *Wine* dataset (classification, 13 features).

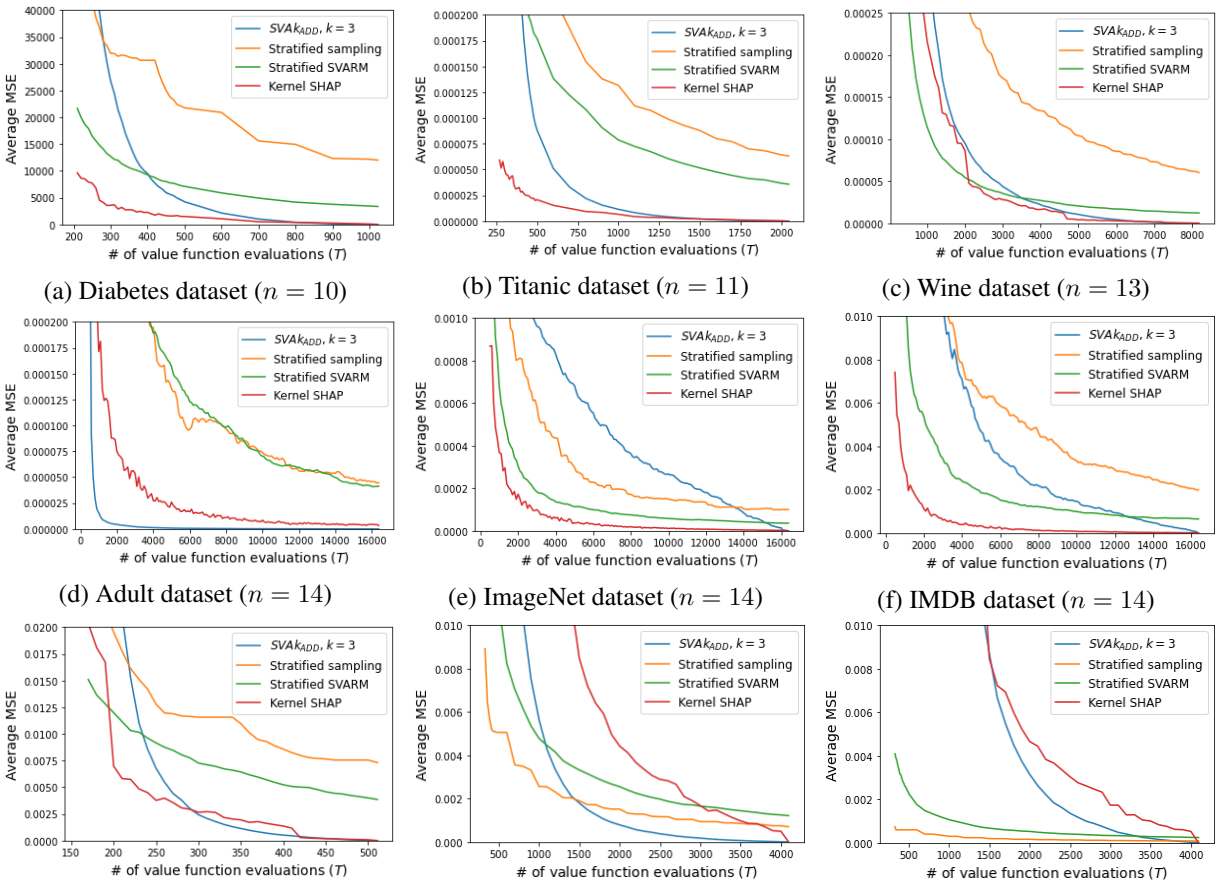
On the contrary, local feature attribution (Lundberg & Lee, 2017) measures each feature's impact on the prediction of a fixed model for a given datapoint. While the predicted value can directly be used as the worth of a feature coalition for regression, the predicted class probability is required instead 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), *ImageNet* (classification, 14 features), and *IMDB* natural language sentiment (regression, 14 features) data.

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 *Breast cancer* (9 features), *Big Five* (12 features), and *FIFA 21* (12 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 (i.e., for different k-additive models).

Figure 2 presents the obtained results for all datasets and for $k \in \{1, 2, 3, 4\}$. Note that 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



(a) Diabetes dataset ($n = 10$) (b) Titanic dataset ($n = 11$) (c) Wine dataset ($n = 13$)

(d) Adult dataset ($n = 14$) (e) ImageNet dataset ($n = 14$) (f) IMDB dataset ($n = 14$)

(g) Breast Cancer dataset ($n = 9$) (h) Big Five dataset ($n = 12$) (i) FIFA dataset ($n = 12$)

Figure 3: MSE of $SVAk_{\text{ADD}}$ and competing methods averaged over 100 repetitions in dependence of available budget T. Datasets stem from various explanation types: global (a)-(c), local (d)-(f), and unsupervised (g)-(i) with differing player numbers n.

behavior for low k, specifically $k = 1$, 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. A similar observation can be made for the 2-additive model in both global and local tasks. Although in FIFA dataset the 2-additive model rapidly converges to the exact Shapley values, for the other ones a higher number of samples are needed until convergence. 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 parmeterization with $k = 1$ or $k = 2$ after a few samples. By comparing $k = 3$ and $k = 4$ variants, the choice of $k = 3$ appears preferable as it results in quicker decreasing error curves.

5.3. Comparison with Existing Approximation 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 & 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 Appendix D for results including *Permutation sampling* (Castro et al., 2009) and the 2-additive model.

First to mention is that $SVAk_{\text{ADD}}$ competes consistently with Stratified SVARM for the best approximation performance across most datasets. Although for a very low number of function evaluations $SVAk_{\text{ADD}}$ achieves an error greater than some other approaches (specially *Stratified SVARM*), 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. The comparison with *KernelSHAP* provides mixed results. For Adult, Big Five and FIFA datasets, *SVA*k_{ADD} converged faster to the exact Shaley values whereas for Titanic and IMDB datasets, *KernelSHAP* achieves a better performance.

6. Conclusion

We proposed with *SVA*k_{ADD} a new algorithm to approximate Shapley values. It falls into the class of approaches that fit a structured surrogate game to the observed value function 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, *SVA*k_{ADD} exhibits competitive results with other existing approaches depending on the considered explanation type, dataset, and available for budget for sampling, 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 begin with in order to obtain a unique solution of the optimization problem, eventually posing a practical limit on k. Adopting further techniques to the sampling procedure within our method, serves as a natural avenue for further research to improve approximation performance. 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. Note that, besides the estimated Shapley values, our proposal could also provide the interaction effects when $k \geq 2$. Although we did not address these parameters, future works can extract the estimated interaction indices and use them to investigate redundant or complementary features. For instance, this could be of interest in practical applications where interaction between features are relevant as for example in disease detection.

References

Balestra, C., Huber, F., Mayr, A., and Müller, E. Unsupervised features ranking via coalitional game theory for categorical data. In *Proceedings of Big Data Analytics and Knowledge Discovery (DaWaK)*, pp. 97–111, 2022.

Bilbao, J., Fernández, J., Jiménez-Losada, A., and López, J. Generating functions for computing power indices efficiently. *Top*, 8:191–213, 2000.

Bordt, S. and von Luxburg, U. From shapley values to generalized additive models and back. In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 709–745, 2023.

Brusa, E., Cibrario, L., Delprete, C., and Di Maggio, L. G. Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring. *Applied Sciences (Switzerland)*, 13(4), 2023. doi: 10.3390/app13042038.

Cai, W., Kordabad, A. B., and Gros, S. Energy management in residential microgrid using model predictive control-based reinforcement learning and Shapley value. *Engineering Applications of Artificial Intelligence*, 119 (January):105793, 2023. doi: 10.1016/j.engappai.2022.105793.

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

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

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

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

Cohen, S. B., Ruppin, E., and Dror, G. Feature selection based on the shapley value. In *Proceedings of the International Joint COnference on Artificial Intelligence (IJCAI)*, pp. 665–670, 2005.

Cohen, S. B., Dror, G., and Ruppin, E. Feature selection via coalitional game theory. *Neural Comput.*, 19(7):1939–1961, 2007.

Covert, I. and Lee, S.-I. Improving kernelshap: Practical shapley value estimation using linear regression. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3457–3465, 2021.

Covert, I., Lundberg, S. M., and Lee, S. Understanding global feature contributions with additive importance measures. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Deng, X. and Papadimitriou, C. H. On the complexity of cooperative solution concepts. *Math. Oper. Res.*, 19(2): 257–266, 1994.

Ebrahimi, M. and Rastegar, M. Towards an interpretable data-driven switch placement model in electric power distribution systems: An explainable artificial intelligence-based approach. *Engineering Applications of Artificial Intelligence*, 129(March 2022):107637, 2024. doi: 10.1016/j.engappai.2023.107637.

Fiestras-Janeiro, M. G., García-Jurado, I., Meca, A., and Mosquera, M. A. Cooperative game theory and inventory management. *European Journal of Operational Research*, 210:459–466, 2011. doi: 10.1016/j.ejor.2010.06.025.

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. SHAP-IQ: unified approximation of any-order shapley interactions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. Kernelshap-iq: Weighted least square optimization for shapley interactions. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Ghorbani, A. and Zou, J. Y. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 2242–2251, 2019.

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

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

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

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

Granot, D., Kuipers, J., and Chopra, S. Cost allocation for a tree network with heterogeneous customers. *Mathematics of Operations Research*, 27(4):647–661, 2002.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gürel, N. M., Li, B., Zhang, C., Spanos, C. J., and Song, D. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endow.*, 12(11):1610–1623, 2019a.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1167–1176, 2019b.

Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. Approximating the shapley value without marginal contributions. In *Proceeedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13246–13255, 2024a.

Kolpaczki, P., Haselbeck, G., and Hüllermeier, E. How much can stratification improve the approximation of shapley values? In *Proceeedings of World Conference on Explainable Artifical Intelligence (xAI)*, pp. 489–512, 2024b.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., and Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9):1–46, 2023. doi: 10.1145/3555803.

Liben-Nowell, D., Sharp, A., Wexler, T., and Woods, K. M. Computing shapley value in supermodular coalitional games. In *Computing and Combinatorics - 18th Annual International Conference COCOON*, pp. 568–579, 2012.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4768–4777, 2017.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., and Rogers, A. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *CoRR*, abs/1306.4265, 2013.

Marcílio, W. E. and Eler, D. M. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 340–347, 2020. doi: 10.1109/SIBGRAPI51738.2020.00053.

Mitchell, R., Cooper, J., Frank, E., and Holmes, G. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022.

Molnar, C. *Interpretable machine learning*. 2021. URL https://christophm.github.io/interpretable-ml-book/.

Murofushi, T. and Soneda, S. Techniques for reading fuzzy measures (iii): interaction index. In *9th fuzzy system symposium*, pp. 693–696, 3 1993.

Nimmy, S. F., Hussain, O. K., Chakrabortty, R. K., Hussain, F. K., and Saberi, M. 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, 2023. doi: 10.1016/j.engappai.2022.105596.

Okhrati, R. and Lipani, A. A multilinear sampling algorithm to estimate shapley values. In *25th International Conference on Pattern Recognition ICPR*, pp. 7992–7999, 2020.

Peleg, B. and Sudhölter, P. *Introduction to the theory of cooperative games*. Springer Science & Business Media, 2 edition, 2007.

Pelegrina, G. D. and Siraj, S. Shapley value-based approaches to explain the quality of predictions by classifiers. *IEEE Transactions on Artificial Intelligence*, pp. 1–15, 2024. doi: 10.1109/TAI.2024.3365082.

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

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

Pelegrina, G. D., Duarte, L. T., and Grabisch, M. Interpreting the contribution of sensors in blind source extraction by means of Shapley values. *IEEE Signal Processing Letters*, 30(1):878–882, 2023b. doi: 10.1109/LSP.2023.3295759.

Pfannschmidt, K., Hüllermeier, E., Held, S., and Neiger, R. 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*, pp. 450–461, 2016.

Rozemberczki, B. and Sarkar, R. The shapley value of classifiers in ensemble games. In *The 30th ACM International Conference on Information and Knowledge Management CIKM*, pp. 1558–1567, 2021.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. The shapley value in machine learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence IJCAI*, pp. 5572–5579, 2022.

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

van Campen, T., Hamers, H., Husslage, B., and Lindelauf, R. 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, 2018.

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

9

SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification

Author Contribution Statement

The author alone developed the idea, algorithm, and analysis. Maximilian Muschalik implemented all experiments and conducted them with some support in the design by Fabian Fumagalli and the author. The author implemented the proposed algorithm and majorly wrote the technical parts, while Fabian Fumagalli and Maximilian Muschalik drafted the introductory sections. The visualizations were done by Maximilian Muschalik. Maximilian Muschalik, Fabian Fumagalli, and the author revised the paper with proofreading of Barbara Hammer and Eyke Hüllermeier.

Supplementary Material

An appendix to the paper is provided in Appendix C.

SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification

Patrick Kolpaczki
Paderborn University

Maximilian Muschalik
University of Munich (LMU)
Munich Center for Machine Learning

Fabian Fumagalli
CITEC
Bielefeld University

Barbara Hammer
CITEC
Bielefeld University

Eyke Hüllermeier
University of Munich (LMU)
Munich Center for Machine Learning

Abstract

Addressing the limitations of individual attribution scores via the Shapley value (SV), the field of explainable AI (XAI) has recently explored intricate interactions of features or data points. In particular, extensions of the SV, such as the Shapley Interaction Index (SII), have been proposed as a measure to still benefit from the axiomatic basis of the SV. However, similar to the SV, their exact computation remains computationally prohibitive. Hence, we propose with SVARM-IQ a sampling-based approach to efficiently approximate Shapley-based interaction indices of any order. SVARM-IQ can be applied to a broad class of interaction indices, including the SII, by leveraging a novel stratified representation. We provide non-asymptotic theoretical guarantees on its approximation quality and empirically demonstrate that SVARM-IQ achieves state-of-the-art estimation results in practical XAI scenarios on different model classes and application domains.

1 INTRODUCTION

Interpreting black box machine learning (ML) models via feature attribution scores is a widely applied technique in the field of explainable AI (XAI) (Adadi and Berrada, 2018; Covert et al., 2021; Chen et al., 2023). However, in real-world applications, such as genomics (Wright et al., 2016) or tasks involving natural language (Tsang et al., 2020), isolated features are less meaningful. In fact, it was shown that, in the presence of strong feature correlation or higher order interactions, feature attribution scores are not sufficient to capture the reasoning of a trained ML model (Wright et al., 2016; Slack et al., 2020; Sundararajan and Najmi, 2020; Kumar et al., 2020, 2021). As a remedy, *feature interactions* extend feature attributions to arbitrary groups of features (see Figure 1).

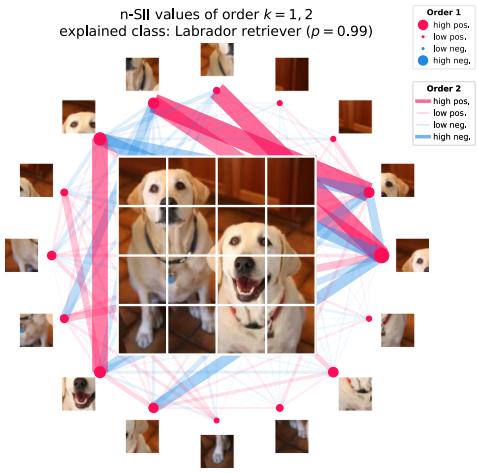


Figure 1: By dividing an ImageNet picture into multiple patches, attribution scores for single patches and interactions scores for pairs aid explaining a vision transformer.

A prevalent approach to define feature attributions is based on the Shapley value (SV) (Shapley, 1953), an axiomatic concept from cooperative game theory

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

that fairly distributes the payout achieved by a group among its members. Extensions of the SV to Shapley-based interaction indices, i.e., interaction indices that reduce to the SV for single players, have been proposed (Grabisch and Roubens, 1999; Bordt and von Luxburg, 2023; Sundararajan et al., 2020; Tsai et al., 2023). Yet, the exact computation of the SV and Shapley-based interactions without further assumptions on the ML model quickly becomes infeasible due to its exponential complexity (Deng and Papadimitriou, 1994).

In this work, we present *SVARM Interaction Quantification* (SVARM-IQ), a novel approximation technique for a broad class of interaction indices, including Shapley-based interactions, which is applicable to any cooperative game. SVARM-IQ extends Stratified SVARM (Kolpaczki et al., 2023) to any-order interactions by introducing a novel representation of interaction indices through stratification.

Contribution. Our core contributions include:

1. *SVARM-IQ* (Section 3): A model-agnostic approximation algorithm for estimating Shapley-based interaction scores of any order through leveraging a *stratified* representation.
2. *Theoretical Analysis* (Section 4): We prove, under mild assumptions, that SVARM-IQ is unbiased and provide bounds on the approximation error.
3. *Application* (Section 5): An open-source implementation[1] and empirical evaluation demonstrating SVARM-IQ's superior approximation quality over state-of-the-art techniques.

Related work. In cooperative game theory, Shapley-based interactions, as an extension to the SV, were first proposed with the Shapley-Interaction index (SII) (Grabisch and Roubens, 1999). Besides the SII, the Shapley-Taylor Interaction index (STI) (Sundararajan et al., 2020) and Faithful Shapley-Interaction index (FSI) (Tsai et al., 2023) were introduced, which, in contrast to the SII, directly require the efficiency axiom. Beyond Shapley-based interaction indices, extensions of the Banzhaf value were studied by Hammer and Holzman (1992). In ML, limitations of feature attribution scores have been discussed in Wright et al. (2016), Sundararajan and Najmi (2020), and Kumar et al. (2020, 2021) among others. Model-specific interaction measures have been proposed for neural networks (Tsang et al., 2018; Singh et al., 2019; Janizek et al., 2021). Model-agnostic measures were introduced via functional decomposition (Hooker, 2004, 2007) in (Lou et al., 2013; Molnar et al., 2019; Lengerich et al., 2020; Hiabu et al., 2023). Applications include complex language (Murdoch et al., 2018) and image classification (Tsang et al., 2020) models, as well as application domains, such as gene interactions (Wright et al., 2016). Besides pure explanation purposes, e.g. understanding sentiment predictions from NLP models (Fumagalli et al., 2023), Chu and Chan (2020) leveraged the SII to improve feature selection for tree classifiers.

Approximation techniques for the SV have been proposed via permutation sampling (Castro et al., 2009), which has been extended to the SII and STI (Sundararajan et al., 2020; Tsai et al., 2023). For the SV, Castro et al. (2017) demonstrated the impact of stratification on approximation performance. Alternatively, the SV can be represented as a solution to a least squares problem (Charnes et al., 1988), which was exploited for approximation (Lundberg and Lee, 2017; Covert and Lee, 2021) and extended to FSI (Tsai et al., 2023). Recent work proposed a model-agnostic sampling-based approach (Fumagalli et al., 2023) for Shapley-based interactions, which was further linked to Covert and Lee (2021). On the model-specific side Muschalik et al. (2024) extended the polynomial-time exact computation of the SV for local feature importance in decision trees (Lundberg et al., 2020) to the SII. While permutation-based approaches are restricted to update single estimates, Kolpaczki et al. (2023) proposed wit Stratified SVARM a novel approach for the SV that is capable of updating all estimates using only a single value function call.

2 SHAPLEY-BASED INTERACTION INDICES

In the following, we are interested in properties of a *cooperative game*, that is a tuple $(\mathcal{N}, \nu)$ containing a *player set* $\mathcal{N} = \{1, \dots, n\}$ with $n \in \mathbb{N}$ players and a *value function* $\nu : 2^{\mathcal{N}} \to \mathbb{R}$ mapping each subset $S \subseteq \mathcal{N}$ of players, also called coalition, to a real-valued number $\nu(S)$. In the field of XAI, the value function typically represents a specific *model behavior* (Covert et al., 2021), such as the prediction of an instance or the dataset loss. The player set represents the entities whose attribution will be determined, e.g., the contribution of features to a prediction or the dataset loss. To determine the worth of individual players, the Shapley value (SV) (Shapley, 1953) can be expressed as a weighted average over marginal contributions.

Definition 2.1 (Shapley Value (Shapley, 1953))**.** *The SV is*

$$\phi_i = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n\binom{n-1}{|S|}} \Delta_i(S),$$

where $i \in \mathcal{N}$ *and* $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$.

[1] https://github.com/kolpaczki/svarm-iq

The SV is provably the unique attribution measure that fulfills the following axioms: linearity (linear combinations of value functions yield linear combinations of attribution), dummy (players that do not impact the worth of any coalition receive zero attribution), symmetry (two players contributing equally to all coalitions receive the same attribution), and efficiency (the sum of of all players' attributions equals the worth of all players) (Shapley, 1953). In many ML related applications, however, the attribution via the SV is limited in the presence of strong feature correlation or higher order interaction (Slack et al., 2020; Sundararajan and Najmi, 2020; Kumar et al., 2020, 2021). It is therefore necessary to study *interactions between players* in cooperative games. The SV is a weighted average of marginal contributions Δ_i of single players, and a natural extension to pairs of players is

$$\Delta_{i,j}(S) := \nu(S \cup \{i,j\}) - \nu(S) - \Delta_i(S) - \Delta_j(S)$$

for $S \subseteq \mathcal{N} \setminus \{i,j\}$. Generalizing this recursion to higher order interactions yields the following definition.

Definition 2.2 (Discrete Derivative (Fujimoto et al., 2006))**.** *For $K \subseteq \mathcal{N}$, the* K-derivative *of ν at $S \subseteq \mathcal{N} \setminus K$ is*

$$\Delta_K(S) := \sum_{W \subseteq K} (-1)^{|K|-|W|} \cdot \nu(S \cup W).$$

The Shapley interaction index (SII) was the first axiomatic extension of the SV to higher order interaction (Grabisch and Roubens, 1999). It can be represented as a weighted average of discrete derivatives.

Definition 2.3 (Shapley Interaction Index (Grabisch and Roubens, 1999))**.** *The SII of $K \subseteq \mathcal{N}$ is defined as*

$$I_K^{SII} = \sum_{S \subseteq \mathcal{N} \setminus K} \frac{1}{(n - |K| + 1)\binom{n-|K|}{|S|}} \Delta_K(S).$$

Cardinal Interaction Indices. Besides the SII, the Shapley-Taylor interaction index (STI) (Sundararajan et al., 2020) and Faithful Shapley interaction index (FSI) (Tsai et al., 2023) have been proposed as extensions of the SV to interactions. More general, the SII can be viewed as a particular instance of a broad class of interaction indices, known as cardinal interaction indices (CIIs) (Fujimoto et al., 2006), which are defined as a weighted average over discrete derivatives:

$$I_K = \sum_{S \subseteq \mathcal{N} \setminus K} \lambda_{k,|S|} \Delta_K(S)$$

with weights $\lambda_{k,|S|}$. In particular, every interaction index satisfying the (generalized) linearity, symmetry and dummy axiom, e.g., SII, STI and FSI, can be represented as a CII (Grabisch and Roubens, 1999). Beyond Shapley-based interaction indices, CIIs also include other interaction indices, such as a generalized Banzhaf value (Hammer and Holzman, 1992). In Section 3, we propose a unified approximation that applies to any CII. For details about other CIIs and their specific weights, we refer to Appendix B.

The SII is the provably unique interaction index that fulfills the (generalized) linearity, symmetry and dummy axiom, as well as a novel recursive axiom that links higher order interactions to lower order interactions (Grabisch and Roubens, 1999). For interaction indices it is also possible to define a generalized efficiency condition, i.e. that $\sum_{K \subseteq \mathcal{N}, |K| \leq k_{\max}} I_K = \nu(\mathcal{N})$ for a maximum interaction order $k_{\max}$. In ML applications, this condition ensures that the sum of contributions equals the model behavior of $\mathcal{N}$, such as the prediction of an instance. The SII scores can be aggregated to fulfill efficiency, which yield the n-Shapley values (n-SII) (Bordt and von Luxburg, 2023). Furthermore, other variants, such as STI and FSI, extend the SV to interactions by directly requiring an efficiency axiom. In contrast to the SV, however, a unique index is only obtained by imposing further conditions. Similar to the SV, whose computation is NP-hard (Deng and Papadimitriou, 1994), the weighted sum of discrete derivatives requires 2^n model evaluations, necessitating approximation techniques.

2.1 Approximations of Shapley-based Interaction Scores

Different approximation techniques have been proposed to overcome the computational complexity of Shapley-based interaction indices, which extend on existing techniques for the SV.

Permutation Sampling. For the SV, permutation sampling (Castro et al., 2009) was proposed, where the SV is represented as an average over randomly drawn permutations of the player set. For each drawn permutation, the algorithm successively adds players to the subset, starting from the empty set using the given order. By comparing the evaluations successively, the marginal contributions are used to update the estimates. Extensions of permutation sampling have been proposed for the SII (Tsai et al., 2023) and STI (Sundararajan et al., 2020). For the SII, only interactions that appear in a consecutive order in the permutation can be updated, resulting in very few updates per permutation. For the STI, all interaction scores can be updated with a single permutation, however, the computational complexity increases, as the discrete derivatives have to be computed for every subset, resulting

in an increase by a factor of 2^k per interaction.

Kernel-based Approximation. Besides the weighted average, the SV also admits a representation as a solution to a constrained weighted least square problem (Charnes et al., 1988). This optimization problem requires again 2^n model evaluations. However, it was proposed to approximate the optimization problem through sampling and solve the resulting optimization problem explicitly, which is known as KernelSHAP (Lundberg and Lee, 2017). An extension of kernel-based approximation was proposed for FSI (Tsai et al., 2023), but it remains open, whether this approach can be generalized to other indices, while its theoretical properties are unknown.

Unbiased KernelSHAP and SHAP-IQ. Unbiased KernelSHAP (Covert and Lee, 2021) constitutes a variant of KernelSHAP to approximate the SV, which yields stronger theoretical results, including an unbiased estimate. While this approach is motivated through a kernel-based approximation, it was shown that it is possible to simplify the calculation to a sampling-based approach (Fumagalli et al., 2023). Using the sampling-based approach, SHAP-IQ (Fumagalli et al., 2023) extends Unbiased KernelSHAP to general interaction indices.

2.2 Stratified Approximation for the SV

Stratification partitions a population into distinct subpopulations, known as strata, where sampling is then separately executed for each stratum. If the strata are chosen as homogeneous groups with lower variability, stratified sampling yields a better approximation. First proposed for the SV by Maleki et al. (2013), it was shown empirically that stratification by coalition size can improve the approximation (Castro et al., 2017), while recent work extended it by more sophisticated techniques (Burgess and Chapman, 2021). With Stratified SVARM, Kolpaczki et al. (2023) proposed an approach that abstains from sampling marginal contributions. Instead, it samples coalitions to leverage its novel representation of the SV, which splits the marginal contributions into two coalitions and stratifies them by size. This allows one to assign each sampled coalition to one stratum per player, thus efficiently computing SV estimates for all players simultaneously. Hence in contrast to permutation sampling, Stratified SVARM reaches a new level of efficiency as all estimates are updated using a single model evaluation. In comparison to KernelSHAP, it is well understood theoretically and shows significant performance improvements compared to Unbiased KernelSHAP (Kolpaczki et al., 2023). In the following, we extend Stratified SVARM to Shapley-based interaction indices, and even general CIIs.

3 SVARM-IQ: A STRATIFIED APPROACH

Since the practical infeasibility of computing the CII incentivizes its approximation as a remedy, we formally state our considered approximation problem under the fixed-budget setting in Section 3.1. We continue by introducing our stratified representation of the CII in Section 3.2, which stands at the core of our new method *SVARM-IQ* presented in Section 3.3.

3.1 Approximation Problem

Given a cooperative game $(\mathcal{N}, \nu)$, an order $k \geq 2$, a budget $B \in \mathbb{N}$, and the weights $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$, with $\mathcal{L}_k := \{0, \ldots, n-k\}$ specifying the desired CII, the goal is to approximate all the latent but unknown CII I_K with $K \in \mathcal{N}_k := \{S \subseteq \mathcal{N} \mid |S| = k\}$ precisely. The budget B is the number of coalition evaluations or in other words accesses to ν that the approximation algorithm is allowed to perform. It captures a time or resource constraint on the computation and is justified by the fact that the access to ν frequently imposes a bottleneck on the runtime due to costly inference, manipulation of data, or even retraining of models. We denote by $\hat{I}_K$ the algorithm's estimate of I_K. Since we consider randomized algorithms, returning stochastic estimates, the approximation quality of an estimate $\hat{I}_K$ is judged by the following two commonly used measures that are to be minimized: First, the mean squared error (MSE) of any set K: $\mathbb{E}\left[(\hat{I}_K - I_K)^2\right]$, and second, a bound on the probability $\mathbb{P}(|\hat{I}_K - I_K| \geq \varepsilon) \leq \delta$ to exceed a threshold $\varepsilon > 0$, commonly known as a (ϵ, δ)-approximation.

3.2 Stratified Representation

Our sampling-based approximation algorithm SVARM-IQ leverages a novel stratified representation of the CII. For the remainder, we stick to the general notion of the CII of any fixed order $k \geq 2$. The concrete interaction type to be approximated can be specified by the weights $\lambda_{k,\ell}$. We stratify the CII I_K by coalition size and split the discrete derivatives $\Delta_K(S)$ into multiple strata to obtain:

$$I_K = \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot I_{K,\ell}^W.$$

with strata terms for all $W \subseteq K$ and $\ell \in \mathcal{L}_k$:

$$I_{K,\ell}^W := \frac{1}{\binom{n-k}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W). \tag{1}$$

This representation is a generalization of the SV representation utilized by Stratified SVARM (Kolpaczki et al., 2023) as it extends from the SV to the CII. Since each stratum contains $\binom{n-k}{\ell}$ many coalitions, $I_{K,\ell}^W$ is a uniform average of all eligible coalition worths and hence we obtain its estimate $\hat{I}_{K,\ell}^W$ by taking the sample-mean of evaluated coalitions belonging to that particular stratum. Further, we can express any CII by means of the strata $I_{K,\ell}^W$ trough manipulating their weighting according to the weights $\lambda_{k,\ell}$. Subsequently, the aggregation of the strata estimates, mimicking our representation, yields the desired CII estimate:

$$\hat{I}_K = \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \hat{I}_{K,\ell}^W.$$

Further, we demonstrate the popular special case of SII between pairs, i.e., $k = 2$, in Appendix D.

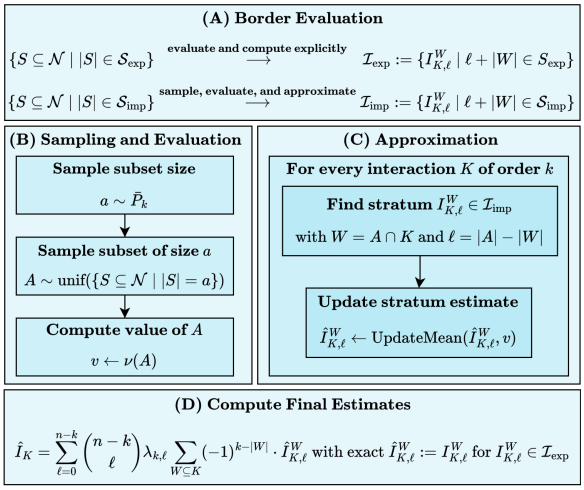


Figure 2: Schematic overview of SVARM-IQ.

3.3 SVARM-IQ

Instead of naively sampling coalitions separately from each of the $2^k\binom{n}{k}(n-k+1)$ many strata, we propose with SVARM-IQ a more sophisticated mechanism, similar to Kolpaczki et al. (2023), which leverages the stratified representation of the CII.

Update Mechanism. SVARM-IQ (given in Algorithm 1 and Figure 2) updates for a single sampled coalition $A \subseteq \mathcal{N}$ one strata estimate of each of the $\binom{n}{k}$ many considered subsets K. This is made feasible by the observation that any coalition A belongs into exactly one stratum associated with I_K. This is in the spirit of the *maximum sample reuse* principle, employed previously for the Banzhaf value (Wang and Jia, 2023) and the SV (Kolpaczki et al., 2023) with the underlying motivation that each seen observation should be utilized to update all interaction estimates. To be more precise, for each $K \in \mathcal{N}_k$ we update

$$\hat{I}_{K,\ell}^W \text{ with } W = A \cap K \text{ and } \ell = |A| - |W|. \quad (2)$$

Notably, our sampling ensures that for *every interaction* $A \setminus K \sim \text{unif}(\{S \subseteq \mathcal{N} \setminus K \mid |S| = \ell\})$, as the probability of $A \setminus K$ conditioned on $W = A \cap K$ and $\ell = |A| - |A \cap K|$ is uniform. This is required as $\hat{I}_{K,\ell}^W$ is a uniform average, cf. Eq. (1), and allows to update estimates for every interaction by sampling a single subset A. Considering the limited budget B, this update rule elicits information from ν in a more "budget-efficient" manner, since it contributes to $\binom{n}{k}$ many estimates with only a single evaluation. To guide the sampling, we first draw in each time step b a coalition size a_b from a probability distribution P_k over the eligible sizes, and draw then A_b uniformly at random among all coalitions of size a_b. We store the evaluated worth $\nu(A_b)$ in order to reuse it for all estimate updates, one for each K. This is done by calling `UpdateMean` (see Appendix C), which sets the associated estimate $\hat{I}_{K,\ell}^W$ to the new average, taking the sampled worth $\nu(A_b)$ and the number of so far observed samples $c_{K,\ell}^W$ of that particular estimate into account. We set P_k to be the uniform distribution over all sizes, i.e., $P_k = \text{unif}(0, n)$. A specifically tailored distribution for $k = 2$ allows us to express sharper theoretical results in Section 4.

Border Sizes. Further, we enhance our approach by transferring a technique, introduced by Fumagalli et al. (2023). We observe that for very low and very high s only a few coalitions of size s exist, $\binom{n}{s}$ many to be precise. Thus, evaluating all these coalitions and calculating the associated strata $I_{K,\ell}^W$ explicitly upfront saves budget, as it avoids duplicates, i.e., coalitions sampled multiple times. Given the budget B and the probability distribution over sizes P_k, we determine a set of subset sizes $\mathcal{S}_{\text{exp}} = \{0, \ldots, s_{\text{exp}}, n - s_{\text{exp}}, \ldots, n\}$, for which the expected number of samples exceeds the number of coalitions of each subset size. Consequently, we evaluate *all coalitions* of sizes in $\mathcal{S}_{\text{exp}}$, i.e., $S \subseteq \mathcal{N}$ with $|S| \in \mathcal{S}_{\text{exp}}$. From the remaining sizes $\mathcal{S}_{\text{imp}} := \{s_{\text{exp}} + 1, \ldots, n - s_{\text{exp}} - 1\}$, we sample coalitions. This split allows to compute all strata

$$\mathcal{I}_{\text{exp}} := \{I_{K,\ell}^W \mid \ell + |W| \in S_{\text{exp}}\}$$

explicitly, which follows from Eq. (2) and $\ell + |W| = |A|$. The remaining strata

$$\mathcal{I}_{\text{imp}} := \{I_{K,\ell}^W \mid \ell + |W| \in \mathcal{S}_{\text{imp}}\}$$

are *approximated* with $\hat{I}_{K,\ell}^W$ by sampling coalitions. The procedure to determine $\mathcal{S}_{\text{exp}}$ and $\mathcal{S}_{\text{imp}}$, named `ComputeBorders` (see Appendix C), is applied before

Algorithm 1 SVARM-IQ

Input: $(\mathcal{N}, \nu)$, $B \in \mathbb{N}$, $k \in \{1, \dots, n\}$, $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$
$\hat{I}^W_{K,\ell}, c^W_{K,\ell} \leftarrow 0 \; \forall K \in \mathcal{N}_k, \ell \in \mathcal{L}_k, W \subseteq K$
$\mathcal{S}_{\text{exp}}, \mathcal{S}_{\text{imp}} \leftarrow$ ComputeBorders
$\bar{B} \leftarrow B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s}$
for $b = 1, \dots, \bar{B}$ **do**
  Draw size $a_b \in \mathcal{S}_{\text{imp}} \sim \bar{P}_k$
  Draw A_b from $\{S \subseteq \mathcal{N} \mid |S| = a_b\}$ u.a.r.
  $v_b \leftarrow \nu(A_b)$ ▷ store coalition worth
  for $K \in \mathcal{N}_k$ **do**
    $W \leftarrow A_b \cap K$ ▷ get stratum set
    $\ell \leftarrow a_b - |W|$ ▷ get stratum size
    $\hat{I}^W_{K,\ell} \leftarrow$ UpdateMean$(\hat{I}^W_{K,\ell}, c^W_{K,\ell}, v_b)$
    $c^W_{K,\ell} \leftarrow c^W_{K,\ell} + 1$ ▷ increment counter
  end for
end for
$\hat{I}_k \leftarrow \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \hat{I}^W_{K,\ell} \; \forall K \in \mathcal{N}_k$
Output: $\hat{I}_K$ for all $K \in \mathcal{N}_k$

the sampling loop in Algorithm 1. Hence, SVARM-IQ enters its sampling loop with a leftover budget of $\bar{B} := B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s}$, and repeatedly applies the update mechanism. The distribution P_k is altered to $\bar{P}_k$ by setting $\bar{P}_k(s) = 0$ for all $s \in \mathcal{S}_{\text{exp}}$ and upscaling all entries $s \in \mathcal{S}_{\text{imp}}$ such that they sum up to 1. Note that this technique yields exact CII values for $B = 2^n$.

Approximating Multiple Orders and Indices. SVARM-IQ is not restricted to approximate only one specific order k at the time. Quite to the contrary, it can be extended to maintain strata estimates $\hat{I}^W_{K,\ell}$ for multiple orders, which are then simultaneously updated within the sampling loop without imposing further budget costs. The aggregation to interaction estimates $\hat{I}_K$ is then carried out for each considered subset K separately. Note that this also entails the SV, i.e., $k = 1$, thus allowing one to approximate attribution and interaction simultaneously. Since the stratification allows to combine the strata to any CII, SVARM-IQ can approximate multiple CII's at the same time, notably without even the need to specify them during sampling. This can be realized by specifying multiple weighting sequences $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$, one for each CII of interest, and performing the final estimate computation $\hat{I}_K$ for each type. Note that this comes without incurring any additional budget cost.

4 THEORETICAL RESULTS

In the following, we present the results of our theoretical analysis for SVARM-IQ. All proofs are deferred to Appendix E. In order to make the analysis feasible, a natural assumption is to observe at least one sample for each approximated stratum $I^W_{K,\ell} \in \mathcal{I}_{\text{imp}}$. We realize this requirement algorithmically only for the remainder of this chapter by executing a WarmUp procedure (see Appendix 3) between ComputeBorders and SVARM-IQ's sampling loop. For each $I^W_{K,\ell} \in \mathcal{I}_{\text{imp}}$ it samples a coalition $A \subseteq \mathcal{N} \setminus K$ of size ℓ and sets $\hat{I}^W_{K,\ell}$ to $\nu(A \cup W)$. Hence, SVARM-IQ enters its sampling loop with a leftover budget of $\tilde{B} := B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s} - |\mathcal{I}_{\text{imp}}|$. We automatically set $s_{\text{exp}} \geq 1$, which consumes only $2n + 2$ evaluations. Hence for $n = 3$, all strata are already explicitly calculated. Since ComputeBorders evaluates then at least all coalitions of size $s \in \{0, 1, n-1, n\}$, the initial distribution P_k over sizes has support $\{2, \dots, n-2\}$. For $k \geq 3$, this allows us to specify P_k to be the uniform distribution:

$$P_k(s) := \frac{1}{n-3} \text{ for all } s \in \{2, \dots, n-2\}.$$

Further for the remainder of the analysis, we use a specifically tailored distribution in the case of $k = 2$:

$$P_2(s) := \begin{cases} \frac{\beta_n}{s(s-1)} & \text{if } s \leq \frac{n-1}{2} \\ \frac{\beta_n}{(n-s)(n-s-1)} & \text{if } s \geq \frac{n}{2} \end{cases}$$

with $\beta_n = \frac{n^2-2n}{2(n^2-4n+2)}$ for even $n \geq 4$ and $\beta_n = \frac{n-1}{2(n-3)}$ for odd $n \geq 5$. This allows us to express sharper bounds in comparison to the uniform distribution.

Notation and assumptions. We introduce some notation, coming in helpful in expressing our results legibly. For any $w \in \{0, \dots, k\}$ we denote by $\mathcal{L}^w_k := \{\ell \in \mathcal{L}_k \mid \ell + w \in \mathcal{S}_{\text{imp}}\}$. For any $K \in \mathcal{N}_k$ and $\ell \in \mathcal{L}_k$ let $A_{K,\ell}$ be a random coalition with distribution $\mathbb{P}(A_{K,\ell} = S) = \binom{n-k}{\ell}^{-1}$ for all $S \subseteq \mathcal{N} \setminus K$ with $|S| = \ell$. For any $W \subseteq K$ we denote the *stratum variance* by $\sigma^2_{K,\ell,W} := \mathbb{V}[\nu(A_{K,\ell} \cup W)]$ and the *stratum range* by $r_{K,\ell,W} := \max_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W) - \min_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W)$. For a comprehensive overview of the used notation, we refer to Appendix A. As our only assumptions, we demand $n \geq 4$ and the budget to be large enough to execute ComputeBorders, WarmUp , and the sampling loop for one iteration, i.e., $\tilde{B} > 0$.

Unbiasedness, Variance, and MSE. We begin by showing that SVARM-IQ's estimates are unbiased, which is not only desirable but will also turn out useful shortly after in our analysis.

Theorem 4.1. *SVARM-IQ's CII estimates are unbiased for all $K \in \mathcal{N}_k$, i.e., $\mathbb{E}[\hat{I}_K] = I_K$.*

The unbiasedness enables us to reduce the MSE of any $\hat{I}_K$ to its variance. In fact, the bias-variance decomposition states that $\mathbb{E}[(\hat{I}_K - I_K)^2] = (\mathbb{E}[\hat{I}_K] - I_K)^2 +$

$\mathbb{V}[\hat{I}_K]$. Hence, a variance analysis of the obtained estimates suffices to bound the MSE. The variance of $\hat{I}_K$ is tightly linked to the number of samples SVARM-IQ collects for each stratum estimate $\hat{I}^W_{K,\ell}$. At this point, we distinguish in our analysis between $k = 2$ and $k \geq 3$ to obtain sharper bounds for the former case facilitated by our carefully designed probability distribution P_2 over coalition sizes. To keep the presented results concise, we introduce $\gamma_k := 2(n-1)^2$ for $k = 2$ and $\gamma_k := n^{k-1}(n-k+1)^2$ for all $k \geq 3$. This stems from the aforementioned difference in precision on the lower bound of collected samples.

Theorem 4.2. *For any $K \in \mathcal{N}_k$ the variance of the CII estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by*

$$\mathbb{V}\left[\hat{I}_K\right] \leq \frac{\gamma_k}{\widetilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W}.$$

Note that our efforts in optimizing the analysis for $k = 2$ reduced the bound by a factor of $\frac{n}{2}$ in comparison to substituting k with 2 in our bound for the general case. This is caused by the severe increase in complexity when trying to give a lower bound for the number of samples each stratum receives. Although our approach allows one to obtain a sharper bound for special cases as $k = 3$ or $k = 4$ with a similarly dedicated analysis, we abstain from doing so as we prioritize a concise presentation of our results.

Corollary 4.3. *For any $K \in \mathcal{N}_k$, the MSE of $\hat{I}_K$ returned by SVARM-IQ is bounded by $\mathbb{E}[(\hat{I}_K - I_K)^2] \leq$*

$$\frac{\gamma_k}{\widetilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W}.$$

We state this result more explicitly for the frequently considered interaction type: the SII for pairs of players i and j. In this case our bound boils down to

$$\mathbb{E}\left[\left(\hat{I}^{\text{SII}}_{i,j} - I^{\text{SII}}_{i,j}\right)^2\right] \leq \frac{2}{\widetilde{B}} \sum_{W \subseteq \{i,j\}} \sum_{\ell \in \mathcal{L}_2^{|W|}} \sigma^2_{i,j,\ell,W}.$$

The simplicity achieved by this result supports a straightforward and natural interpretation. The MSE bound of each SII estimate is inversely proportional to the available budget for the sampling loop and each stratum variance contributes equally to its growth.

We *intentionally abstain* from expressing our bounds in asymptotic notation w.r.t. B and n only, as it would not do justice to the motivation behind employing stratification. The performance of SVARM-IQ is based on lower strata variances (and also strata ranges) compared to the whole population of all coalition values within the powerset of $\mathcal{N}$. This improvement can not be reflected adequately by the asymptotics in which the variances vanish to constants.

(ϵ, δ)-Approximation. Combining Theorem 4.2 with Chebyshev's inequality immediately yields a bound on the probability that the absolute error of a fixed $\hat{I}_K$ exceeds some $\varepsilon > 0$ given the budget at hand.

Corollary 4.4. *For any $K \in \mathcal{N}_k$, the absolute error of $\hat{I}_K$ returned by SVARM-IQ exceeds some fixed ε with probability of at most $\mathbb{P}(|\hat{I}_K - I_K| \geq \varepsilon) \leq$*

$$\frac{\gamma_k}{\varepsilon^2 \widetilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W}.$$

One can easily rearrange the terms to find the minimum budget required to obtain $\mathbb{P}(|\hat{I}_K - I_K| \leq \varepsilon) \geq 1 - \delta$ for a given $\delta > 0$. Note that this bound still depends on the unknown strata variances. Further, we provide another bound in Theorem 4.5 (see Appendix E.4), resulting from a slightly more laborious usage of Hoeffding's inequality which takes the strata ranges into account. To the best of our knowledge there exists no theoretical analysis for permutation sampling of CIIs. SHAP-IQ is like wise unbiased, but its theoretical analysis (Theorem 4.3) Fumagalli et al. (2023) does not provide such detail for fixed n and k.

5 EXPERIMENTS

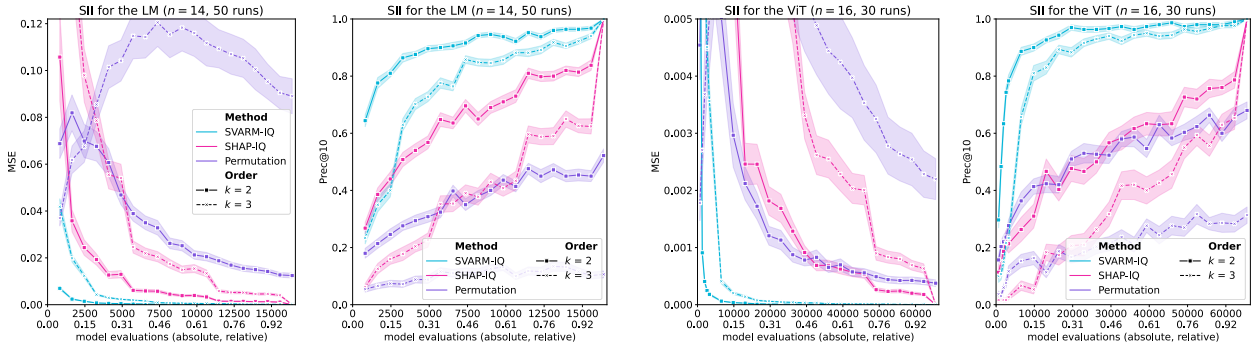
We empirically evaluate SVARM-IQ's approximation quality in different XAI application scenarios and compare it with current state-of-the-art baselines.

Baselines. In the case of estimating SII and STI scores, we compare SVARM-IQ to SHAP-IQ (Fumagalli et al., 2023) and permutation sampling (Sundararajan et al., 2020; Tsai et al., 2023). For FSI, we compare against the kernel-based regression approach (Tsai et al., 2023) instead of permutation sampling.

Table 1: Overview of the XAI tasks and models used

Task	Model ID	Removal Strategy	n	$\mathcal{Y}$
LM	DistilBert	Token Removal	14	$[-1, 1]$
ViT	ViT-32-384	Token Removal	9,16	$[0, 1]$
CNN	ResNet18	Superpixel Marginalization	14	$[0, 1]$

Explanation Tasks. Similar to Fumagalli et al. (2023) and Tsai et al. (2023), we evaluate the approximation algorithms based on different real-world ML models and classical XAI scenarios (cf. Table 1). First, we compute interaction scores to explain a sentiment analysis *language model* (LM), which is a fine-tuned version of `DistilBert` (Sanh et al., 2019) on the



(a) MSE and Prec@10 for the LM

(b) MSE and Prec@10 for the ViT

Figure 3: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines for estimating order $k = 2, 3$ SII on the LM (a; $n = 14$) and the ViT (b; $n = 16$). Shaded bands represent the standard error over 50, respectively 30 runs.

IMDB (Maas et al., 2011) dataset. Second, we investigate two types of image classification models, which were pre-trained on ImageNet (Deng et al., 2009). We explain a *vision transformer* (ViT), (Dosovitskiy et al., 2021), and a `ResNet18` *convolutional neural network* (CNN) (He et al., 2016a). The ViT operates on patches of 32 times 32 pixels and is abbreviated with `ViT-32-384`. The `torch` versions of the LM, ViT, and the CNN are retrieved from Wolf et al. (2020) and Paszke et al. (2017). For further descriptions on the models and feature removal strategies aligned with Covert et al. (2021), we refer to Appendix F.

Measuring Performance. To assess the performance of the different approximation algorithms, we measure the mean squared error averaged over all $K \in \mathcal{N}_k$ (MSE; lower is better) and the precision at ten (Prec@10; higher is better) of the estimated interaction scores compared to pre-computed ground-truth values (GTV). Prec@10 measures the ratio of correctly identifying the ten highest (absolute) interaction values. The GTV for each run are computed exhaustively with 2^n queries to the black box models. All results are averaged over multiple independent runs.

Approximation Quality for SII. We compare SVARM-IQ against permutation sampling and SHAP-IQ at the LM and ViT explanation tasks for approximating all SII values of order $k = 2$ and $k = 3$ in Figure 3. Across both considered measures, MSE and Prec@10, SVARM-IQ demonstrates superior approximation quality. Noteworthy is SVARM-IQ's steep increase in approximation quality in the earlier budget range allowing applications with limited computational resources. Based on our theoretical findings, we assume the stratification by size in combination with the splitting of discrete derivatives to be the cause for

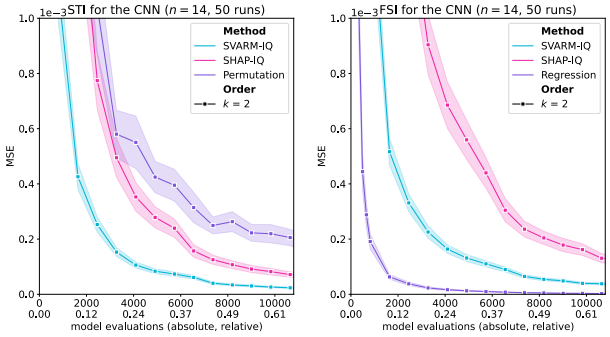


Figure 4: Comparison of SVARM-IQ and baselines for STI (left) and FSI (right) on the CNN. Shaded bands represent the standard error over 50 runs.

the observed behavior. Most plausibly coalitions of the same size and sharing a predetermined set, as encompassed by each stratum $I^W_{K,\ell}$, vary less in their worth than the whole population of coalitions. Consequently, the associated variance $\sigma^2_{K,\ell,W}$ is considerably lower, leading to faster convergence of the estimate $\hat{I}^W_{K,\ell}$.

Example Use-Case of n-SII Values. Precise estimates allow to construct high-quality n-SII scores as proposed by Bordt and von Luxburg (2023). Figure 1 illustrates how n-SII scores can be used to explain the ViT with 16 patches for an image of two correctly classified Labradors. All individual patches receive positive attribution scores ($k = 1$) of varying degree, leading practitioners to assume that patches with similar attribution are of equal importance. However, enhancing the explanation with second order interactions ($k = 2$), reveals how the interplay between patches containing complementing facial parts, like the eyes and the mouth, strongly influences the model's prediction towards the correct class label. On the contrary,

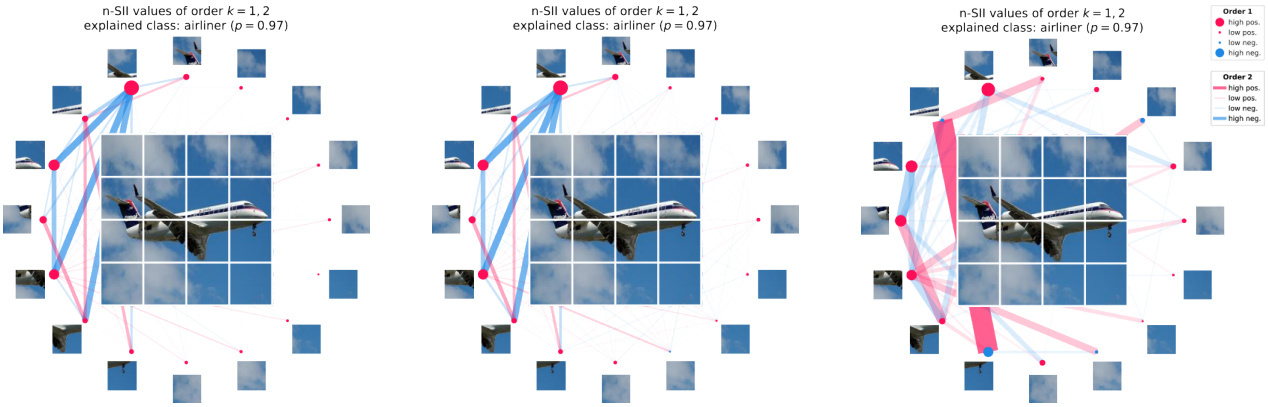


Figure 5: Comparison of ground-truth n-SII values of order $k = 1$ and $k = 2$ for the predicted class probability of a ViT for an ImageNet picture sliced into a grid of $n = 16$ patches (left) against n-SII values estimated by SVARM-IQ (center) and permutation sampling (right). The exact computation requires 65,536 model evaluations while the budget of both approximators is limited by 5000, making up only 7.6% of the space to sample.

tiles depicting the same parts, e.g. those containing eyes, show negative interaction, allowing to conclude that the addition of one in the presence of the other is on average far less impactful than their individual contribution. Solely observing the monotony of the individual scores would have arguably led to overlook this insight. We describe this further in Appendix G.2.

Estimating FSI and STI. Further, we compute different CIIs of a fixed order with SVARM-IQ and consistently achieve high approximation quality. We summarize the results on the CNN in Figure 4. For STI, SVARM-IQ, again, outperforms both sampling-based baselines. The kernel-based regression estimator, which is only applicable to the FSI index, yields lower approximation errors than SVARM-IQ. Similar to SV estimation through KernelSHAP (Lundberg and Lee, 2017), this highlights the expressive power of the least-squares representation available for FSI.

Instance-wise comparison. Lastly, we compare in Figure 5 SVARM-IQ's n-SII estimates of order $k = 1$ and $k = 2$ against those of permutation sampling and the ground truth for single a single instance. The ground truth interaction is computed upfront for the predicted class probability of the ViT for a specified image sliced into a grid of 16 patches, and both approximation algorithms are executed for a single run with a budget of 5000 model evaluations, thus consuming only 7.6% of the budget necessary to compute GTV exactly. The estimates obtained by SVARM-IQ show barely any visible difference to the human eye. In fact, SVARM-IQ's approximation replicates the ground truth with only a fraction of the number of model evaluations that are necessary for its exact computation. Hence, it significantly lowers the computational burden for precise explanations. On the contrary, permutation sampling yields estimated importance and interaction scores which are afflicted with evident imprecision. This lack in approximation quality has the potential to cause misguiding explanations. More comparisons are shown in Appendix G.2.

6 CONCLUSION

We proposed SVARM-IQ, a new sampling-based approximation algorithm for interaction indices based on a stratified representation to maximize budget efficiency. SVARM-IQ is capable of approximating all types and orders of cardinal interactions simultaneously, including the popular SII. Consequently, as the special case of SVs is also entailed, this facilitates the approximation of feature importance and interaction simultaneously, thus offering an enriched explanation. Besides proving theoretical results, we empirically demonstrated SVARM-IQ's advantage against current state-of-the-art baselines. Its model-agnostic nature and domain-independence allow practitioners to obtain high-quality interaction scores for various entity types such as features or data points.

Limitations and Future Work. Due to SVARM-IQ's stratification, the number of maintained strata estimates grows exponentially with the interaction order k. This space complexity poses a challenge for large interaction orders. As a pragmatic remedy, future work may consider the approximation of interaction scores for a smaller number of sets or a coarser stratification by size. Lastly, it still remains unclear whether the performance of the kernel-based regression estimator available for FSI and the SV can be transferred to other types of CIIs like the SII or STI indices.

Acknowledgements

This reserach was supported supported by the research training group Dataninja (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia. Maximilan Muschalik and Fabian Fumagalli gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282.

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

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 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 709–745. PMLR.

Burgess, M. A. and Chapman, A. C. (2021). Approximating the shapley value using stratified empirical bernstein sampling. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, (IJCAI) 2021*, pages 73–81. ijcai.org.

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.

Chao, M. T. and Strawderman, W. E. (1972). Negative Moments of Positive Random Variables. *Journal of the American Statistical Association*, 67(338):429–431.

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*, volume 11 of *Advanced Studies in Theoretical and Applied Econometrics*, page 123–133. Springer Netherlands.

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

Chu, C. and Chan, D. P. K. (2020). Feature selection using approximated high-order interaction components of the shapley value for boosted tree classifier. *IEEE Access*, 8:112742–112750.

Covert, I. and Lee, S. (2021). Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, volume 130 of *Proceedings of Machine Learning Research*, pages 3457–3465. PMLR.

Covert, I., Lundberg, S. M., and Lee, S. (2021). Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research*, 22:209:1–209:90.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255. IEEE Computer Society.

Deng, X. and Papadimitriou, C. H. (1994). On the Complexity of Cooperative Solution Concepts. *Mathematics of Operations Research*, 19(2):257–266.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations (ICLR 2021)*. OpenReview.net.

Fujimoto, K., Kojadinovic, I., and Marichal, J. (2006). Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99.

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. (2023). SHAP-IQ: Unified Approximation of any-order Shapley Interactions. *CoRR*, abs/2303.01179.

Grabisch, M. and Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565.

Hammer, P. L. and Holzman, R. (1992). Approximations of pseudo-Boolean functions; applications to game theory. *ZOR Mathematical Methods of Operations Research*, 36(1):3–21.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778. IEEE Computer Society.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778. IEEE Computer Society.

Hiabu, M., Meyer, J. T., and Wright, M. N. (2023). Unifying local and global model explanations by functional decomposition of low dimensional structures. In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 7040–7060. PMLR.

Hooker, G. (2004). Discovering additive structure in black box functions. In Kim, W., Kohavi, R., Gehrke, J., and DuMouchel, W., editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2004)*, pages 575–580. ACM.

Hooker, G. (2007). Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732.

Janizek, J. D., Sturmfels, P., and Lee, S. (2021). Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *Journal of Machine Learning Research*, 22:104:1–104:54.

Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. (2023). Approximating the Shapley Value without Marginal Contributions. *CoRR*, abs/2302.00736.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. A. (2021). Shapley Residuals: Quantifying the limits of the Shapley value for explanations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021 (NeurIPS 2021)*, pages 26598–26608.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR.

Lengerich, B. J., Tan, S., Chang, C., Hooker, G., and Caruana, R. (2020). Purifying Interaction Effects with the Functional ANOVA: An Efficient Algorithm for Recovering Identifiable Additive Models. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2402–2412. PMLR.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013)*, pages 623–631. ACM.

Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NeurIPS 2017)*, pages 4765–4774.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

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.

Molnar, C., Casalicchio, G., and Bischl, B. (2019). Quantifying Model Complexity via Functional Decomposition for Better Post-hoc Interpretability. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*, volume Communications in Computer and Information Science, pages 193–204. Springer, Cham.

Murdoch, W. J., Liu, P. J., and Yu, B. (2018). Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *6th International Conference on Learning Representations (ICLR 2018)*.

Muschalik, M., Fumagalli, F., , Hüllermeier, E., and Hammer, B. (2024). Beyond TreeSHAP: Efficient Computation of Any-Order Shapley Interactions for Tree Ensembles. *CoRR*, abs/2401.12069.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *Workshop at Conference on Neural Information Processing Systems (NeurIPS 2017)*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*.

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

Singh, C., Murdoch, W. J., and Yu, B. (2019). Hierarchical interpretations for neural network predictions. In *7th International Conference on Learning Representations (ICLR 2019)*.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES 2020)*, pages 180–186. ACM.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9259–9268. PMLR.

Sundararajan, M. and Najmi, A. (2020). The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR.

Tsai, C., Yeh, C., and Ravikumar, P. (2023). FaithShap: The Faithful Shapley Interaction Index. *Journal of Machine Learning Research*, 24(94):1–42.

Tsang, M., Cheng, D., Liu, H., Feng, X., Zhou, E., and Liu, Y. (2020). Feature Interaction Interpretability: A Case for Explaining Ad-Recommendation Systems via Neural Interaction Detection. In *8th International Conference on Learning Representations (ICLR 2020)*.

Tsang, M., Cheng, D., and Liu, Y. (2018). Detecting Statistical Interactions from Neural Network Weights. In *6th International Conference on Learning Representations (ICLR 2018)*.

Wang, J. T. and Jia, R. (2023). Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 6388–6421. PMLR.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020)*, pages 38–45. Association for Computational Linguistics.

Wright, M. N., Ziegler, A., and König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, 17:145.

10 Identifying Top-k Players in Cooperative Games via Shapley Bandits

Author Contribution Statement

The author alone developed the idea, algorithms, analysis, and experiment design. The author also implemented the algorithms and experiments, and conducted the experiments on his own. Both coauthors supported the writing based on a draft written by the author.

Identifying Top-k Players in Cooperative Games via Shapley Bandits

Patrick Kolpaczki[1], Viktor Bengs[1] and Eyke Hüllermeier[2]

[1]Paderborn University, Germany
[2]University of Munich (LMU), Germany

Abstract

The usefulness of cooperative game theory and key concepts like the Shapley value, which measures the contribution of individual players to the overall performance of a coalition, has been demonstrated in various applications. Due to the computational effort growing exponentially with the number of participants in a game, several methods have been proposed to approximate Shapley values. Yet, in many applications, only the order of players according to their Shapley values is important, or maybe the set of the k best players, but not the values themselves. In this paper, we consider the problem of identifying the k players in a cooperative game with the highest Shapley values and denote it as the *Top-k Shapley* problem. By viewing the marginal contributions of a player as a random variable, we establish a connection between cooperative games and multi-armed bandits, which in turn allows us to reduce Top-k Shapley to the multiple arms identification problem. We call the resulting bandits problem *Shapley bandits*. Besides adopting existing algorithms for multiple arms identifications, we propose the *Border Uncertainty Sampling* algorithm (BUS) and provide empirical evidence for its superiority over state-of-the-art algorithms.

Keywords

Shapley value, Cooperative games, Multi-armed bandit, Multiple arms identification

1. Introduction

The formal notion of a cooperative game, in which players can form coalitions to accomplish a certain task, is a versatile concept with countless practical applications. Consider, for example, the cooperation of municipalities in infrastructure projects, with the goal to reduce costs by sharing and allocating available resources. In the context of (supervised) machine learning, individual features can be seen as players and feature subsets as coalitions — the task here is to train a model with high predictive performance [1, 2].

An interesting question in the context of cooperative games concerns the importance or contribution of an individual player: How to distribute the collective benefit of a coalition among the individual players? A connection to explainable AI can be drawn by interpreting features in a machine learning model as players and the predictive performance as the collective benefit such that the portion allocated to each feature can be seen as its importance for the model. Independent of the considered application, cooperative game theory has proposed different solution concepts, with the Shapley value as the arguably most popular one [3]. The Shapley

LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany
patrick.kolpaczki@upb.de (P. Kolpaczki); viktor.bengs@upb.de (V. Bengs); eyke@ifi.lmu.de (E. Hüllermeier)

value assigns to each player a weighted average of all its marginal contributions, where we understand by a marginal contribution of a player the increase in the worth of a coalition when adding that player. The popularity of the Shapley value arises from the fact that it can be derived axiomatically by demanding desirable properties that one would expect from a fair distribution [3]. It has found its usage in a broad range of fields, from identifying influential members in terrorist networks [4, 5] to finding important neurons in artificial neural networks [6].

An inherent drawback of the Shapley value is the huge computational effort caused by the exponentially (in the number of players) growing number of marginal contributions — one per coalition — to be averaged over. As a consequence, brute force approaches quickly become infeasible for even only a few dozens of players. Several approximation methods have been proposed [6, 7, 8] to tackle this difficulty, all of them sharing the same idea of calculating mean estimates for randomly sampled marginal contributions uniformly for all players. Further, theoretical guarantees for approximation methods have been shown under mild assumptions [7, 8].

While these approximations show partially satisfying results in empirical studies, it seems to be rarely mentioned that in many applications the true objective is not to obtain precise Shapley value estimates for all players, but to identify a certain number of k players with the highest Shapley values (even though most works are indirectly aiming for that). For example, security agencies are more interested in identifying the most threatening members in terrorist networks, or the good performance of a machine learning model is oftentimes largely driven only by the most valuable features. Needless to say, one could tackle this problem näively by just pointing at the k players with highest Shapley value estimates obtained by traditional approximation algorithms. However, this approach would involve sampling steps to approximate the Shapley values of players for which one can already be certain that these are at the top or bottom of the ranking in terms of the Shapley values. In such cases, on the other hand, it makes sense to sample marginal contributions for players lying in the ”middle” of the ranking in order to separate as quickly as possible the set of k-best players from the rest with a certain degree of certainty, although this might involve sacrificing precision of estimates for those players who are likely to be at the top or bottom of the ranking.

Similar considerations have already been made in the field of multi-armed bandit (MAB) problems [9], which is a class of online learning problems, where an agent needs to choose one arm (choice alternative) among a given set of arms (choice alternatives) in the course of a sequential decision process to achieve a specific target. In the stochastic variant of the MAB problem, each arm is associated with an unknown reward distribution and choosing a specific arm results in obtaining a stochastic reward generated by the chosen arm's unknown reward distribution. Many of the targets considered therefore revolve around identifying a specific partial ranking with respect to the (unknown) means of the arms reward distributions as quickly as possible. One particular target is to find the k arms having the highest mean, known as the multiple arms identifications problem [10], for which a number of algorithmic solutions are already available [10, 11, 12, 13, 14]. In this paper, we show how to trace the Top-k Shapley problem back to the multiple arms identifications problem, so that state-of-the-art solution methods for the latter problem can be efficiently used for the former. In addition, we propose a new method that performs even superior in numerical experiments.

2. Preliminaries

Before introducing our proposed problem formally in Section 3, we revisit in the following cooperative games and the Shapley value, as well as the problem of multiple arms identification in multi-armed bandit problems.

2.1. Cooperative Games and the Shapley Value

A *cooperative game* is characterized by a pair (N, ν) containing a set of *players* $N = \{p_1, \ldots, p_n\}$ and a *value function* $\nu : \mathcal{P}(N) \to \mathbb{R}$, where $\nu(\emptyset) = 0$ by definition. The players can form *coalitions* $S \subseteq N$ and obtain a combined benefit given by $\nu(S)$ which is called the *worth* of S. For the question of how to distribute the worth $\nu(N)$ of the *grand coalition* N to the individual n many players, the *Shapley value* [3] forms a payoff distribution allocating to each player p_i the value

$$\phi_i(\nu) = \sum_{S \subseteq N \setminus \{p_i\}} \frac{1}{n\binom{n-1}{|S|}} \cdot (\nu(S \cup \{p_i\}) - \nu(S)).$$

For simplicity, we write ϕ_i whenever it is clear to which value function ν we refer. The difference in worth $\nu(S \cup \{p_i\}) - \nu(S)$ is called p_i's *marginal contribution* given S. The Shapley value can be derived axiomatically, as it is provably the only solution concept fulfilling simultaneously the following properties [3], which one would intuitively demand from a fair distribution:

- Efficiency: the worth of N is partitioned over all players, i.e., $\nu(N) = \sum_{p_i \in N} \phi_i$,
- Symmetry: if two players p_i and p_j cannot be distinguished by their marginal contributions, i.e., $\nu(S \cup \{p_i\}) = \nu(S \cup \{p_j\})$ for all $S \subseteq N$ not containing p_i or p_j, then $\phi_i = \phi_j$,
- Additivity: if ν is a sum of two value functions ν_1 and ν_2, i.e., $\nu = \nu_1 + \nu_2$, then $\phi_i(\nu) = \phi_i(\nu_1) + \phi_i(\nu_2)$,
- Dummy element: if a player p_i has constant marginal contribution $\nu(\{p_i\})$ for all coalitions, i.e., $\nu(S \cup \{p_i\}) = \nu(S) + \nu(\{p_i\})$ for all $S \subseteq N \setminus \{p_i\}$, then $\phi_i = \nu(\{i\})$.

2.2. Multiple Arms Identification

A *multi-armed bandit* problem is specified by a set of *arms* $\mathcal{A} = \{a_1, \ldots, a_n\}$ each arm a_i of which is endowed with an unknown distribution ζ_i having mean μ_i. In each discrete *time step* t, the learner can *pull* an arm a_i of its choice, meaning that it retrieves a random sample $X_i^t \sim \zeta_i$ drawn independently conditioned on the history of the previous time steps. The arms can be ordered (not necessarily uniquely) via a permutation $\pi : [n] \to [n]$ such that $\mu_{\pi(1)} \geq \ldots \geq \mu_{\pi(n)}$, where we define $[n] := \{1, \ldots, n\}$. Given a number $k \in [n]$, the objective of the learner in the *multiple arms identification* problem is to identify the top-k arms $a_{\pi(1)}, \ldots, a_{\pi(k)}$. In the literature there are two prevalent learning frameworks for this objective, namely the *fixed budget* setting and the *fixed confidence* setting. In the former, a number of time steps T (the *budget*) is given beforehand, which once exhausted requires the learner to return its guess about the top-k arms, with its performance being measured by the probability of returning a correct output. On the contrary, the learner is judged in the latter by the number

of time steps needed in order to identify the top-k arms with probability at least $1 - \delta$ for a given $\delta \in (0, 1]$.

3. Problem Statement

The *Top-k Shapley* problem is given by a cooperative game (N, ν) in which accesses to the value function ν are costly. Although ν is known (in the sense that we can access $\nu(S)$ for all $S \subseteq N$), the Shapley values remain unknown, since it is practically infeasible for a sufficiently large number of players to compute them. The players in N can be ordered (not necessarily uniquely) via a permutation $\pi : [n] \to [n]$ such that $\phi_{\pi(1)} \geq \ldots \geq \phi_{\phi(n)}$. For sake of simplicity, we assume that the there are no ties at the top-k-th position. Given a number $k \in [n]$, the learner's goal is to identify the top-k players $p_{\pi(1)}, \ldots, p_{\pi(k)}$ with highest Shapley values.

Likewise to multiple arms identification, we distinguish between two learning scenarios. One where performance is measured by the probability of the learner successfully identifying the top k players after a given number T of accesses to ν that the learner is allowed to make (fixed budget scenario). The other focusing on a minimal number of accesses to ν in order to guarantee a successful identification with a probability of at least $1 - \delta$ for a given $\delta \in (0, 1]$ (fixed confidence scenario). Due to page restrictions we focus only on the fixed budget setting.

4. Reduction to Multiple Arms Identification

Given a cooperative game (N, ν), the marginal contribution $\nu(S \cup \{p_i\}) - \nu(S)$ of each player p_i can be viewed as a discrete random variable X_i if S is drawn randomly from $\mathcal{P}(N \setminus \{p_i\})$. Further, by drawing any S with probability $1/n\binom{n-1}{|S|}$, X_i has mean $\mathbb{E}[X_i] = \phi_i$. Thus, by interpreting a player p_i as an arm a_i within a multi-armed bandit problem, where retrieving a sample of the arm's distribution corresponds to drawing a (independent) sample of X_i, we obtain that the arm's mean μ_i equals the player's Shapley value ϕ_i. Together with the Shapley values, the corresponding arms' means remain unknown to us. With this connection at hand, the reduction to multiple arms identification is complete, as the objective of identifying the top-k players $p_{\pi(1)}, \ldots, p_{\pi(k)}$ with highest Shapley values is equivalent to the task of finding the corresponding k arms $a_{\pi(1)}, \ldots, a_{\pi(k)}$ having highest means. We denote the resulting bandit problem as *Shapley bandits*. This general reduction scheme allows leveraging any algorithm for multiple arms identification to the Top-k Shapley problem without affecting its internal mechanisms. Finally, it should be emphasized that each pull of an arm a_i involves two accesses to the value function ν, one for $\nu(S)$ and the other for $\nu(S \cup \{p_i\})$.

5. Algorithms

We present and analyze in Section 5.1 *Uniform Random Sampling* as a first benchmark algorithm, show in Section 5.2 how to adapt already existing algorithms for multiple arms identification to the top-k Shapley problem at the example of the *Gap-based Exploration* algorithm [14], and propose in Section 5.3 with *Border Uncertainty Sampling* a new algorithm that can be easily generalized to multiple arms identification.

5.1. Uniform Random Sampling

As an illustrative example of how the approach can be applied we present the *Uniform Random Sampling* algorithm (see Algorithm 1). It is a modification of the *ApproShapley* algorithm in [7] and the *Simple Random Sampling* algorithm in [8], which instead of sampling permutations of players and computing marginal contributions in the sequence in which players in the permutations appear, simply samples a coalition for each player in order to remain faithful to our reduction explained above (cf. Section 4).

For each player p_i a mean estimate $\hat{\phi}_i$ of ϕ_i is kept by URS and at termination the k players with highest estimates are returned. Note how URS does not rely on a budget T or confidence $1-\delta$ to be given, instead it can be run for an arbitrary number of time steps and is therefore applicable for the fixed budget setting as well as the fixed confidence setting. Utilizing the

Algorithm 1: Uniform Random Sampling (URS)

Input: N, ν, k
Initialize: $\hat{\phi}_i \leftarrow 0, t_i \leftarrow 0 \; \forall p_i \in N$
for $t = 1, 2, \ldots$ **do**
  $i \leftarrow (t \mod n) + 1$
  $t_i \leftarrow t_i + 1$
  $\phi_{i,t_i} = \nu(S \cup \{p_i\}) - \nu(S)$ with $S \subseteq N \setminus \{p_i\}$ drawn with probability $1/n\binom{n-1}{|S|}$
  $\hat{\phi}_i \leftarrow \frac{(t_i-1)\hat{\phi}_i + \phi_{i,t_i}}{t_i}$
end
Output: $p_{\hat{\pi}(1)}, \ldots, p_{\hat{\pi}(k)}$ for $\hat{\pi} : [n] \to [n]$ with $\hat{\phi}_{\hat{\pi}(1)} \geq \ldots \geq \hat{\phi}_{\hat{\pi}(n)}$

techniques presented in [8], we can derive performance guarantees for the fixed budget and the fixed confidence setting depending on the variances or ranges of the marginal contributions of each player, stated in the following.

Theorem 1.
Let $\sigma^2 \geq \mathbb{V}[X_i]$ for all $p_i \in N$ and $k \in [n]$, $m \in \mathbb{N}$, $\delta \in (0, 1]$, as well as $\varepsilon_k > 0$ with $\varepsilon_k \leq \phi_{\pi(k)} - \phi_{\pi(k+1)}$. Then, URS identifies the top-k players correctly

- after $2mn$ many accesses to ν with probability at least $1 - {4n\sigma^2}/{\varepsilon_k^2 m}$;
- with probability at least $1-\delta$ after ${8n^2\sigma^2}/{\varepsilon_k^2 \delta}$ many accesses to ν.

The proof is given in Appendix A. The first property becomes a guarantee for the fixed budget scenario by setting m (denoting the number of marginal contributions drawn for each player) to the highest integer fulfilling $2mn \leq T$ for the given budget T. The second property reveals a sampling complexity of ${8n^2\sigma^2}/{\varepsilon_k^2 \delta}$ for the fixed confidence scenario.

Theorem 2.
Let r be an upper bound for the range of X_i for all $p_i \in N$. Further, let $k \in [n]$, $m \in \mathbb{N}$, $\delta \in (0, 1]$, and $0 < \varepsilon_k \leq \phi_{\pi(k)} - \phi_{\pi(k+1)}$. Then, URS identifies the top-k players correctly

- after $2mn$ many accesses to ν with probability at least $1 - 2n \exp\left(-{\varepsilon_k^2 m}/{2r^2}\right)$;

- with probability at least $1-\delta$ after $4nr^2/\varepsilon_k^2 \cdot \log(2n/\delta)$ many accesses to ν.

The proof is given in Appendix B. Again, m is to be interpreted as the number of marginal contributions drawn for each player.

5.2. Gap-based Exploration

At the example of the *Gap-based Exploration* algorithm (Gap-E) [15, 14] we demonstrate how to adapt a multiple arms identification algorithm to the Top-k Shapley problem (see Algorithm 2). Originally, Gap-E was proposed and analyzed for the setting of finding the single arm with highest mean reward in [15], and later slightly modified for the task of finding the top-k arms in [14]. Whenever Gap-E pulls an arm a_i, we replace the random sample by $\nu(S \cup \{i\}) - \nu(S)$ for $S \subseteq N \setminus \{p_i\}$ drawn randomly with probability $1/n\binom{n-1}{|S|}$. Gap-E demands the budget T, a coefficient $c \in \mathbb{R}_{>0}$, and the complexity of the problem $H^{\langle k \rangle}$ as additional parameters to be given, where

$$H^{\langle k \rangle} = \sum_{i=1}^{n} \left(\Delta_i^{\langle k \rangle}\right)^{-2}, \quad \text{and} \quad \Delta_i^{\langle k \rangle} = \begin{cases} \mu_i - \mu_{\pi(k+1)}, & i \in \{\pi(1), \dots, \pi(k)\} \\ \mu_{\pi(k)} - \mu_i, & i \in \{\pi(k+1), \dots, \pi(n)\} \end{cases}.$$

Algorithm 2: Gap-based Exploration (Gap-E)

Input: N, ν, T, c, $H^{\langle k \rangle}$
Initialize: $\hat{\phi}_i \leftarrow 0, t_i \leftarrow 1 \; \forall p_i \in N$
for $i = 1, \dots, n$ **do**
  $\hat{\phi}_i = \nu(S \cup \{p_i\}) - \nu(S)$ with $S \subseteq N \setminus \{p_i\}$ drawn with probability $1/n\binom{n-1}{|S|}$
end
for $t = n+1, \dots, T$ **do**
  Compute $\hat{\pi} : [n] \to [n]$ with $\hat{\phi}_{\hat{\pi}(1)} \geq \dots \geq \hat{\phi}_{\hat{\pi}(n)}$
  $\Delta_i = \begin{cases} \hat{\phi}_i - \hat{\phi}_{\hat{\pi}(k+1)} & i \in \{\hat{\pi}(1), \dots, \hat{\pi}(k)\} \\ \hat{\phi}_{\hat{\pi}(k)} - \hat{\phi}_i & i \in \{\hat{\pi}(k+1), \dots, \hat{\pi}(n)\} \end{cases} \quad \forall p_i \in N$
  $i \leftarrow \arg\max_{j \in [n]} -\Delta_j + c\sqrt{\frac{T}{H^{\langle k \rangle} t_j}}$
  $t_i \leftarrow t_i + 1$
  $\phi_{i,t_i} = \nu(S \cup \{p_i\}) - \nu(S)$ with $S \subseteq N \setminus \{p_i\}$ drawn with probability $1/n\binom{n-1}{|S|}$
  $\hat{\phi}_i \leftarrow \frac{(t_i - 1)\hat{\phi}_i + \phi_{i,t_i}}{t_i}$
end
Output: $p_{\hat{\pi}(1)}, \dots, p_{\hat{\pi}(k)}$ for $\hat{\pi} : [n] \to [n]$ with $\hat{\phi}_{\hat{\pi}(1)} \geq \dots \geq \hat{\phi}_{\hat{\pi}(n)}$

5.3. Border Uncertainty Sampling

Next, we propose a new algorithm (cf. Algorithm 3) called *Border Uncertainty Sampling* (BUS) without providing theoretical guarantees. In similar fashion to Gap-E a measure of (un-)certainty

whether a player p_i belongs to the top-k players or not is at the heart of BUS. However, the gaps involved in the measure of (un-)certainty are calculated in a slightly different manner, namely as the absolute distance to the average of the k-th and $(k+1)$-th highest mean estimates $\hat{\phi}_{\hat{\pi}(k)}$ and $\hat{\phi}_{\hat{\pi}(k+1)}$. Next, BUS chooses to draw a sample for the player p_i that minimizes its gap times the number of samples BUS has already drawn for it, i.e., $\Delta_i \cdot t_i$. The intuition behind this measure of certainty is that for players with larger gap Δ_i we are more certain to tell whether it belongs to the top-k players or not. Likewise, a larger number t_i of samples drawn indicates a higher precision of the estimate $\hat{\phi}_i$. Thus, BUS selects the player p_i with highest uncertainty. As with URS, a clear advantage of BUS over Gap-E is that no additional parameters like the time budget for instance are required, allowing it to be terminated at any time step.

Algorithm 3: Border Uncertainty Sampling (BUS)

Input: N, ν, k
Initialize: $\hat{\phi}_i \leftarrow 0, t_i \leftarrow 1 \; \forall p_i \in N$
for $i = 1, \ldots, n$ **do**
  $\hat{\phi}_i = \nu(S \cup \{p_i\}) - \nu(S)$ with $S \subseteq N \setminus \{p_i\}$ drawn with probability $1/n\binom{n-1}{|S|}$
end
for $t = n+1, \ldots$ **do**
  Compute $\hat{\pi} : [n] \to [n]$ with $\hat{\phi}_{\hat{\pi}(1)} \geq \ldots \geq \hat{\phi}_{\hat{\pi}(n)}$
  $\hat{\phi}^* \leftarrow \frac{\hat{\phi}_{\hat{\pi}(k)} + \hat{\phi}_{\hat{\pi}(k+1)}}{2}$
  $\Delta_i \leftarrow |\hat{\phi}_i - \hat{\phi}^*| \; \forall p_i \in N$
  $i \leftarrow \arg\min_{j \in [n]} \Delta_j \cdot t_j$
  $t_i \leftarrow t_i + 1$
  $\phi_{i,t_i} = \nu(S \cup \{p_i\}) - \nu(S)$ with $S \subseteq N \setminus \{p_i\}$ drawn with probability $1/n\binom{n-1}{|S|}$
  $\hat{\phi}_i \leftarrow \frac{(t_i - 1)\hat{\phi}_i + \phi_{i,t_i}}{t_i}$
end
Output: $p_{\hat{\pi}(1)}, \ldots, p_{\hat{\pi}(k)}$ for $\hat{\pi} : [n] \to [n]$ with $\hat{\phi}_{\hat{\pi}(1)} \geq \ldots \geq \hat{\phi}_{\hat{\pi}(n)}$

6. Experiments

In the following we evaluate the algorithms URS, BUS, Gap-E [14], and *Successive Accepts and Rejects* (SAR) [14] modified for the Top-k Shapley problem on synthetic data. For Gap-E we have heuristically set $H^{\langle k \rangle} = 10000$ and $c = 1$. We are interested in the performance curves in dependence of the number of players n, the budget T, and the variance in marginal contributions. Generating random value functions is not suitable for our purpose, as this leads to expensive computations of the corresponding Shapley values. As a remedy, we simulated cooperative games with the following two approaches. First, we consider in Section 6.1 a stochastic setting in which the marginal contributions of each player are sampled from some fixed distributions. And secondly, we simulate in 6.2 a special case of cooperative games called *sum of unanimity*

games for which the computation of Shapley values is fairly straightforward. We show in all figures for each choice of parameters the averaged ratio of correctly identified top-k players gathered from 500 repetitions.

6.1. Stochastic Setting

We substitute the marginal contributions of each player p_i by a random variable $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and set $\mu_i = 0.806 - 0.006i$ for all $p_i \in N$. The results are shown in Figure 1. For all three considered dependencies (budget, number of players, and variances) BUS outperforms the other considered algorithms by a visible margin. The performance of all algorithms improves for increasing budgets and decreasing variances as one would expect, but the impact of the number of players on BUS's and Gap-E's ratio is surprisingly low in the considered ranges.

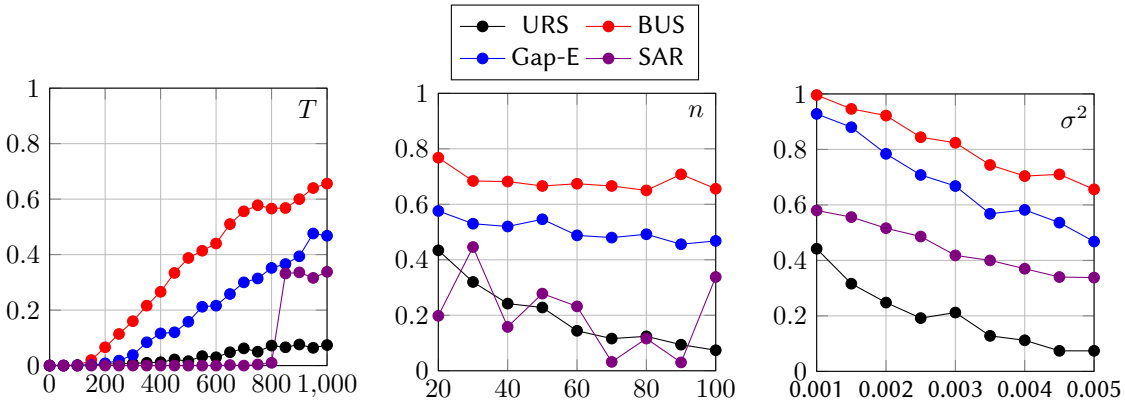


Figure 1: Averaged ratios of correct returned sets under the stochastic setting for $k = 10$. Left: $n = 100$, $\sigma^2 = 0.005$. Center: $T = 1000$, $\sigma^2 = 0.005$. Right: $n = 100$, $T = 1000$.

6.2. Sums of Unanimity Games

In an *unanimity game*, specified by a subset $R \subseteq N$, the value function takes the form of

$$\nu_R(S) = \mathbb{I}\{R \subseteq S\} \text{ for all } S \subseteq N,$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. An unanimity game can be interpreted as a game in which all players contained in R have to agree on cooperating together in order to achieve a benefit of 1. One can construct a *sum-of-unanimity-games game* (SOUG game) by combining multiple unanimity games in a linear combination. More precisely, for a set of coalitions $\mathcal{R} \subseteq \mathcal{P}(N)$ and coefficients $c_R \in \mathbb{R}$ for each $R \in \mathcal{R}$ the value function is given by:

$$\nu(S) = \sum_{R \in \mathcal{R}} c_R \cdot \nu_R(S) \text{ for all } S \subseteq N.$$

The Shapley values of a SOUG game can be calculated in linear time with respect to the number of combined unanimity games and is given for each player p_i by [3]:

$$\phi_i = \sum_{R \in \mathcal{R}: i \in R} \frac{c_R}{|R|}.$$

For our simulations we generate SOUG games by drawing all the key terms uniformly at random within a specific range/domain, respectively. The considered ranges or domains are

- $\{5, 6, \ldots, 50\}$ for the number of combined unanimity games $|\mathcal{R}|$,
- $\{0, 1, \ldots, n\}$ for the size of each $R \in \mathcal{R}$,
- N for the members of each $R \in \mathcal{R}$,
- $[0, 1/|\mathcal{R}|]$ for the coefficient c_R for each $R \in \mathcal{R}$.

The results in Figure 2 show a similar picture as for the stochastic setting, albeit the performance ratios being closer together. BUS still outperforms its competitors Gap-E and SAR, while the benchmark algorithm URS does not perform significantly worse, which indicates the increased challenge that SOUG games pose in comparison to the stochastic setting. In contrast, the number of players has now a more drastic impact.

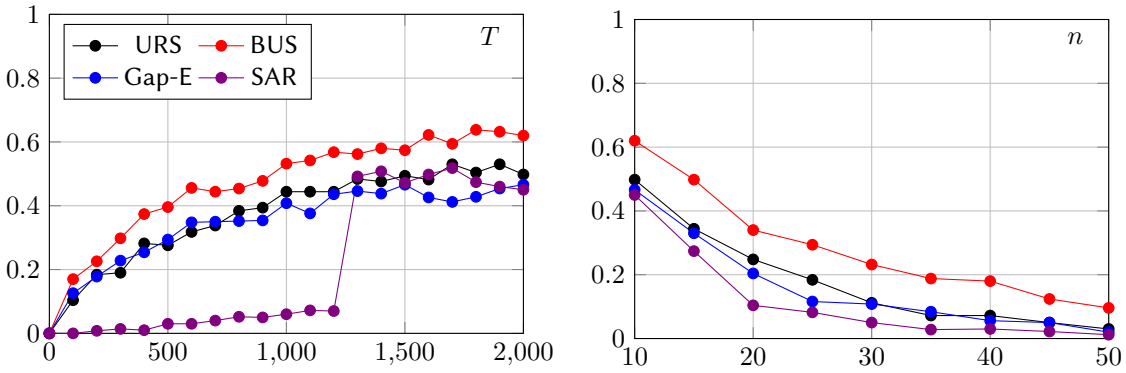


Figure 2: Averaged ratios of correct returned sets for SOUG games for $k = 3$. Left: $n = 10$. Right: T=2000.

7. Conclusion

We have proposed the Top-k Shapley problem, which consists of finding the k players in a cooperative game with the highest Shapley values. Taking a probabilistic view by seeing the marginal contributions of the players as discrete random variables allowed us to draw a connection to multi-armed bandits and reduce the problem to multiple-arms identification, which we have done by successfully adapting known algorithms. We proposed with BUS a new algorithm that is not limited to the use case of identifying top-k Shapley players and gave evidence for its superiority by means of empirical results. Further, it has the advantage of not

needing to know any additional parameters compared to other algorithms for multiple arms identification. For future work, we aim to derive theoretical guarantees, albeit leaving room for modifications open in order to make the analysis feasible.

Acknowledgments

This research was supported by the research training group Dataninja (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] S. B. Cohen, G. Dror, E. Ruppin, Feature selection via coalitional game theory, Neural Comput. 19 (2007) 1939–1961.
[2] K. Pfannschmidt, E. Hüllermeier, S. Held, R. Neiger, Evaluating tests in medical diagnosis: Combining machine learning with game-theoretical concepts, in: IPMU, volume 610, Springer, 2016, pp. 450–461.
[3] L. S. Shapley, 17. A Value for n-Person Games, Princeton University Press, 2016, pp. 307–318.
[4] T. van Campen, H. Hamers, B. Husslage, R. Lindelauf, A new approximation method for the shapley value applied to the WTC 9/11 terrorist attack, Soc. Netw. Anal. Min. 8 (2018) 3:1–3:12.
[5] R. Lindelauf, H. Hamers, B. Husslage, Cooperative game theoretic centrality analysis of terrorist networks: The cases of jemaah islamiyah and al qaeda, Eur. J. Oper. Res. 229 (2013) 230–238.
[6] A. Ghorbani, J. Y. Zou, Neuron shapley: Discovering the responsible neurons, in: Advances in Neural Information Processing Systems 33, 2020.
[7] J. Castro, D. Gómez, J. Tejada, Polynomial calculation of the shapley value based on sampling, Comput. Oper. Res. 36 (2009) 1726–1730.
[8] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, A. Rogers, Bounding the estimation error of sampling-based shapley value approximation with/without stratifying, CoRR abs/1306.4265 (2013).
[9] T. Lattimore, C. Szepesvári, Bandit Algorithms, Cambridge University Press, 2020.
[10] S. Kalyanakrishnan, P. Stone, Efficient selection of multiple bandit arms: Theory and practice, in: Proceedings of the 27th International Conference on Machine Learning, Omnipress, 2010, pp. 511–518.
[11] S. Kalyanakrishnan, A. Tewari, P. Auer, P. Stone, PAC subset selection in stochastic multi-armed bandits, in: Proceedings of the 29th International Conference on Machine Learning, icml.cc / Omnipress, 2012.
[12] S. Chen, T. Lin, I. King, M. R. Lyu, W. Chen, Combinatorial pure exploration of multi-armed bandits, in: Advances in Neural Information Processing Systems 27, 2014, pp. 379–387.
[13] Y. Zhou, X. Chen, J. Li, Optimal PAC multiple arm identification with applications to

crowdsourcing, in: Proceedings of the 31th International Conference on Machine Learning, volume 32, JMLR.org, 2014, pp. 217–225.
[14] S. Bubeck, T. Wang, N. Viswanathan, Multiple identifications in multi-armed bandits, in: Proceedings of the 30th International Conference on Machine Learning, volume 28, JMLR.org, 2013, pp. 258–265.
[15] V. Gabillon, M. Ghavamzadeh, A. Lazaric, S. Bubeck, Multi-bandit best arm identification, in: Advances in Neural Information Processing Systems 24, 2011, pp. 2222–2230.

A. Proof of Theorem 1

For all i and t_i we can view ϕ_{i,t_i} as a discrete random variable with:

$$\begin{aligned}\mathbb{E}[\phi_{i,t_i}] &= \sum_{S\subseteq N\setminus\{i\}} \frac{1}{n\cdot\binom{n-1}{|S|}}\cdot\nu(S\cup\{i\})-\nu(S)\\ &= \phi_i.\end{aligned}$$

Let T_i be the number of times marginal contributions have been drawn for i and $Y_i = \sum_{t_i=1}^{T_i}\phi_{i,t_i}$, thus $\mathbb{E}[Y_i] = T_i\phi_i$ and $\hat{\phi}_i = {}^{Y_i}/_{T_i}$ at the point of termination.

Lemma 3.
Let $\varepsilon_k > 0$ with $\varepsilon_k \leq \phi_{\pi(k)} - \phi_{\pi(k+1)}$. The probability of URS identifying the top-k Shapley players correctly is at least

$$1-\sum_{i=1}^{n}\mathbb{P}\left(|\hat{\phi}_{\pi(i)}-\phi_{\pi(i)}|\geq\frac{\varepsilon_k}{2}\right).$$

Proof:
First, we show that a correct identification of the top-k players by URS implies that all Shapley values are estimated with an absolute error of at most $\frac{\varepsilon_k}{2}$:

$$\begin{aligned}&\bigcup_{i=1}^{k}\bigcup_{j=k+1}^{n}\left\{\hat{\phi}_{\pi(i)}\leq\hat{\phi}_{\pi(j)}\right\}\\ =&\bigcup_{i=1}^{k}\bigcup_{j=k+1}^{n}\left\{\left(\hat{\phi}_{\pi(j)}-\phi_{\pi(j)}\right)+\left(\phi_{\pi(i)}-\hat{\phi}_{\pi(i)}\right)\geq\phi_{\pi(i)}-\phi_{\pi(j)}\right\}\\ \subseteq&\bigcup_{i=1}^{k}\bigcup_{j=k+1}^{n}\left\{\hat{\phi}_{\pi(j)}-\phi_{\pi(j)}\geq\frac{\phi_{\pi(i)}-\phi_{\pi(j)}}{2}\right\}\cup\left\{\phi_{\pi(i)}-\hat{\phi}_{\pi(i)}\geq\frac{\phi_{\pi(i)}-\phi_{\pi(j)}}{2}\right\}\\ \subseteq&\bigcup_{i=1}^{k}\bigcup_{j=k+1}^{n}\left\{|\hat{\phi}_{\pi(j)}-\phi_{\pi(j)}|\geq\frac{\varepsilon_k}{2}\right\}\cup\left\{|\hat{\phi}_{\pi(i)}-\phi_{\pi(i)}|\geq\frac{\varepsilon_k}{2}\right\}\end{aligned}$$

$$= \bigcup_{i=1}^{n} \left\{ |\hat{\phi}_{\pi(i)} - \phi_{\pi(i)}| \geq \frac{\varepsilon_k}{2} \right\}.$$

From which we derive:

$$\mathbb{P}\left(\bigcap_{i=1}^{k} \bigcap_{j=k+1}^{n} \left\{ \hat{\phi}_{\pi(i)} > \hat{\phi}_{\pi(j)} \right\} \right) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(|\hat{\phi}_{\pi(i)} - \phi_{\pi(i)}| \geq \frac{\varepsilon_k}{2} \right). \qquad \square$$

Let $\sigma_i^2 = \mathbb{V}[\phi_{i,t_i}]$ and hence $\mathbb{V}[Y_i] = T_i \sigma_i^2$. Similar to [8], we obtain by using Chebyshev's inequality for all $\varepsilon_k > 0$:

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \frac{\varepsilon_k}{2} \right) \leq \frac{4\sigma_i^2}{\varepsilon_k^2 T_i}.$$

We complete the proof by deriving for $\sigma \geq \sigma_i$ and $m \leq T_i$ for all i with the help of Lemma 3:

$$\mathbb{P}\left(\bigcap_{i=1}^{k} \bigcap_{j=k+1}^{n} \left\{ \hat{\phi}_{\pi(i)} > \hat{\phi}_{\pi(j)} \right\} \right) \geq 1 - \frac{4n\sigma^2}{\varepsilon_k^2 m}.$$

B. Proof of Theorem 2

Let r_i be the range of ϕ_{i,t_i} for all i. Similar to [8], we obtain by using Hoeffding's inequality, for all $\varepsilon_k > 0$:

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \frac{\varepsilon_k}{2} \right) \leq 2 \exp\left(-\frac{\varepsilon_k^2 T_i}{2 r_i^2} \right).$$

We complete the proof by deriving for $r \geq r_i$ and $m \leq T_i$ for all i with the help of Lemma 3:

$$\mathbb{P}\left(\bigcap_{i=1}^{k} \bigcap_{j=k+1}^{n} \left\{ \hat{\phi}_{\pi(i)} > \hat{\phi}_{\pi(j)} \right\} \right) \geq 1 - 2n \exp\left(-\frac{\varepsilon_k^2 M}{2 r^2} \right).$$

11 Antithetic Sampling for Top-k Shapley Identification

Author Contribution Statement

The author developed the idea of the paper together with Tim Nielen and was responsible for the paper project while Tim Nielen majorly contributed to the algorithms and analysis. The author created the visualization. Tim Nielen implemented and conducted all experiments while the experiment design was developed jointly. The paper was written and edited by the author based on a draft for the technical parts by Tim Nielen who further contributed by revising and proofreading the manuscript. Eyke Hüllermeier proofread the paper.

Supplementary Material

An appendix to the paper is provided in Appendix D.

Antithetic Sampling for Top-k Shapley Identification

Patrick Kolpaczki [1 2] **Tim Nielen** [1] **Eyke Hüllermeier** [1 2]

Abstract

Additive feature explanations rely primarily on game-theoretic notions such as the Shapley value by viewing features as cooperating players. The Shapley value's popularity in and outside of explainable AI stems from its axiomatic uniqueness. However, its computational complexity severely limits practicability. Most works investigate the uniform approximation of all features' Shapley values, needlessly consuming samples for insignificant features. In contrast, identifying the k most important features can already be sufficiently insightful and yields the potential to leverage algorithmic opportunities connected to the field of multi-armed bandits. We propose *Comparable Marginal Contributions Sampling* (CMCS), a method for the *top-k identification problem* utilizing a new sampling scheme taking advantage of correlated observations. We conduct experiments to showcase the efficacy of our method in compared to competitive baselines. Our empirical findings reveal that estimation quality for the *approximate-all problem* does not necessarily transfer to *top-k identification* and vice versa.

1. Introduction

The fast-paced development of artificial intelligence poses a double-edged sword. Obviously on one hand, machine learning models have significantly improved in prediction performance, most famously demonstrated by deep learning models. But, on the other hand, their required complexity to exhibit these capabilities comes at a price. Human users face concerning challenges comprehending the decision-making of such models that appear to be increasingly opaque. The field of explainable AI (Vilone & Longo, 2021; Molnar, 2022) offers a simple yet popular approach to regain understanding and shed light onto these black box models by means of *additive feature explanations* (Doumard et al., 2022). Probing a model's behavior to input, this explanation method assigns importance scores to the utilized features. Depending on the explanandum of interest, each score can be interpreted as the feature's impact on the models' prediction for a particular instance or its generalization performance.

The Shapley value (Shapley, 1953) has emerged as a prominent mechanism to assign scores. Taking a game-theoretic perspective, each feature is viewed as a player in a *cooperative game* in which the players can form coalitions and reap a *collective benefit* by solving a task together. For instance, a coalition representing a feature subset can be rewarded with the generalization performance of the to be explained model using only that subset. Posing the omnipresent question of how to divide in equitable manner the collective benefit that all players jointly achieve, reduces the search for feature importance scores to a *fair-division problem*. The Shapley value is the unique solution to fulfill certain desiderata which arguably capture an intuitive notion of fairness (Shapley, 1953). The marginal contributions of a player to all coalitions, denoting the increase in collective benefit when joining a coalition, are taken into a weighted sum by the Shapley value.

It has been extensively applied for *local explanations*, dividing the prediction value (Lundberg & Lee, 2017), and *global explanations* that divide prediction performance (Covert et al., 2020). In addition to providing understanding, other works proposed to utilize it for the selection of machine learning entities such as features (Cohen et al., 2007; Wang et al., 2024), datapoints (Ghorbani & Zou, 2019), neurons in deep neural networks (Ghorbani & Zou, 2020), or base learners in ensembles (Rozemberczki & Sarkar, 2021). We refer to (Rozemberczki et al., 2022) for an overview of its applications in machine learning. Unfortunately, the complexity of the Shapley value poses a serious limitation: its calculation encompasses all coalitions within the exponentially growing power set of players. Hence, the exact computation of the Shapley value is quickly doomed for even moderate feature numbers. Ergo, the research branch of estimating the Shapley value has sparked notable interest, in particular the challenge of precisely approximating the Shapley values of all players known as the *approximate-all problem*.

However, often the exact importance scores just serve as a means to find the most influential features, be it for explana-

[1]LMU Munich [2]Munich Center for Machine Learning. Correspondence to: Patrick Kolpaczki <patrick.kolpaczki@ifi.lmu.de>.

tion or preselection (Cohen et al., 2007; Wang et al., 2024), and are not particularly relevant themselves. Hence, we advocate for the *top-k identification problem* (Kolpaczki et al., 2021) in which an approximation algorithm's goal is to identify the k players with highest Shapley values, without having to return precise estimates. This incentivizes to forego and sacrifice precision of players' estimates for whom reliable predictions of top-k membership already manifest during runtime. Instead, the available samples, reflecting finite computational power at disposal, are better spent on players on the verge of belonging to the top-k in order to speed up the segregation of top-k players from the rest.

Contribution. We propose with *Comparable Marginal Contributions Sampling* (CMCS), Greedy CMCS, and CMCS@K novel top-k identification algorithms for the Shapley value. More specifically, our contributions are:

- We present a new representation of the Shapley value based on an altered notion of marginal contribution and leverage it to develop CMCS. On the theoretical basis of antithetic sampling, we underpin the intuition behind utilizing correlated observations especially for top-k identification.
- Moreover, with Greedy CMCS and CMCS@K we propose multi-armed bandit-inspired enhancements. Our proposed algorithms are model-agnostic and applicable to any cooperative game independent of the domain of interest.
- Lastly, we observe how empirical performance does not directly translate from the approximate-all to the top-k identification problem. Depending on the task, different algorithms are favorable and a conscious choice is advisable.

2. Related Work

The problem of precisely approximating all players' Shapley values has been extensively investigated. Since the Shapley value is a weighted average of a player's marginal contributions, methods that conduct mean estimation form a popular class of approximation algorithms. Most of these sample marginal contributions as performed by *ApproShapley* (Castro et al., 2009). Many variance reduction techniques, that increase the estimates' convergence speed, have been incorporated: stratification (Maleki et al., 2013; O'Brien et al., 2015; Castro et al., 2017; van Campen et al., 2018; Okhrati & Lipani, 2020; Burgess & Chapman, 2021), antithetic sampling (Illés & Kerényi, 2019; Mitchell et al., 2022), and control variates (Goldwasser & Hooker, 2024). Departing from the notion of marginal contributions, other methods view the Shapley value as a composition of coalition values and sample these instead for mean estimation (Covert et al., 2019; Kolpaczki et al., 2024a;b). A different class of methods does not approximate Shapley values directly, but fits a parametrized surrogate game via sampling. As the surrogate game represents the game of interest increasingly more faithful, its own Shapley values become better estimates. Due to the surrogate game's highly restrictive structure these can be obtained in polynomial time. *KernelSHAP* (Lundberg & Lee, 2017) is the most prominent member of this class with succeeding extensions (Covert & Lee, 2021; Pelegrina et al., 2025). See (Chen et al., 2023) for an overview of further methods for feature attribution and specific model classes.

First to consider the top-k identification problem for Shapley values were Narayanam & Narahari (2008) by simply returning the players with the highest estimates effectively computed by *ApproShapley* (Castro et al., 2009). This straightforward reduction of top-k identification to the approximate-all problem can be realized with any approximation algorithm. Kolpaczki et al. (2021) establish a connection to the field of multi-armed bandits (Lattimore & Szepesvári, 2020) and thus open the door to further algorithmic opportunities that top-k identification has to offer. Here, pulling an arm of a slot machine metaphorically captures the draw of a sample from a distribution. Usually, one is interested in maximizing the cumulative random reward obtained from sequentially playing the multi-armed slot machine or finding the arm with highest mean reward. Modeling each player as an arm and its reward distribution to be the player's marginal contributions distributed according to their weights within the Shapley value (Kolpaczki et al., 2021), facilitates the usage of bandit algorithms to find the k distributions with highest mean values which represent the players' Shapely values. The inherent trade-off between constantly collecting information from all arms to avoid falling victim to the estimates' stochasticity and selecting only those players that promise the most information gain to correctly predict top-k membership, constitutes the well-known exploration-exploitation dilemma.

Bandit algorithms such as *Gap-E* (Gabillon et al., 2011) and *Border Uncertainty Sampling (BUS)* (Kolpaczki et al., 2021) tackle it by greedily selecting the next arm to pull as the one that maximizes a selection criterion which combines the uncertainty of top-k membership and its sample number. In contrast *Successive Accepts and Rejects (SAR)* (Bubeck et al., 2013) phase-wise eliminates arms whose top-k membership can be reliably predicted. *SHAP@K* (Kariyappa et al., 2024) employs an alternative greedy selection criterion based on confidence intervals for the players' estimates. In each round, samples are taken from two players, one from the currently predicted top-k and one outside of them, with the highest overlap in confidence intervals. The overlap is interpreted as the likelihood that the pair is mistakenly partitioned and should be swapped instead.

3. The Top-k Identification Problem

We introduce cooperative games and the Shapley value formally in Section 3.1, and briefly after present the widely studied problem of approximating all players' Shapley values in a cooperative game Section 3.2. On that basis, we introduce the problem of identifying the top-k players with the highest Shapley values in Section 3.3 and distinguish it from the former by highlighting decisive differences in performance measures which will prepare our theoretical findings and arising methodological avenues alluded to in Section 4.

3.1. Cooperative Games and the Shapley Value

A cooperative game $(\mathcal{N}, \nu)$ consists of a *player set* $\mathcal{N} = \{1, \dots, n\}$ and a *value function* $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ that maps each subset $S \subseteq \mathcal{N}$ to a real-valued worth. The players in $\mathcal{N}$ can cooperate by forming *coalitions* in order to achieve a goal. A coalition is represented by a subset S of $\mathcal{N}$ that includes exactly all players which join the coalition. The formation of a coalition resolves in the (partial) fulfillment of the goal and a collective benefit $\nu(S)$ disbursed to the coalition which we call the *worth* of that coalition. The empty set has no worth, i.e. $\nu(\emptyset) = 0$. The abstractness of this notion offers a certain versatility in modeling many cooperative scenarios. In the context of feature explanations for example, each player represents a feature and the formation of a coalition is interpreted to express that a model or learner uses only that feature subset and discards those features absent in the coalition. Depending on the desired explanation type, the prediction value for a datapoint of interest or an observed behavior of the model over multiple instances, for example generalization performance on a test set, is commonly taken as the worth of a feature subset.

A central problem revolving around cooperative games is the question of how to split the collective benefit that all players achieve together among them. More precisely, which share ϕ_i of the *grand coalition's* worth $\nu(\mathcal{N})$ should each player $i \in \mathcal{N}$ receive? A common demand is that these payouts ϕ are to be fair and reflect the contribution that each player provides to the fulfillment of the goal. Guided by this rationale, the Shapley value (Shapley, 1953) offers a popular solution by assigning each player i the payoff

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

The difference in worth $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$ is known as *marginal contribution* and reflects the increase in collective benefit that i causes by joining the coalition S. The reason for the Shapley value's popularity lies within its axiomatic justification. It is the unique payoff distribution to simultaneously satisfy the four axioms, symmetry, linearity, efficiency, and dummy player (Shapley, 1953), which capture an intuitive notion of fairness in light of the faced fair division problem. Despite this appeal, the Shapley value comes with a severe drawback. The number of coalition values contained in its summation grows exponentially w.r.t. the number of players n in the game. In fact, its exact calculation is provably NP-hard (Deng & Papadimitriou, 1994) if no further assumption on the structure of ν is made, and as a consequence, the Shapley value becomes practically intractable for datasets with even medium-sized feature numbers. This issue necessitates the precise estimation of Shapley values to provide accurate explanations.

3.2. Approximating all Shapley Values

Within the *approximate-all problem*, the objective of an approximation algorithm $\mathcal{A}$ is to precisely estimate the Shapley values $\phi = (\phi_1, \dots, \phi_n)$ of all players by means of estimates $\hat{\phi} = (\hat{\phi}_1, \dots, \hat{\phi}_n)$ for a given cooperative game $(\mathcal{N}, \nu)$. We consider the *fixed-budget* setting in which the number of times $\mathcal{A}$ can access ν to evaluate the worth $\nu(S)$ of a coalition S of its choice is limited by a budget $T \in \mathbb{N}$. Thus, $\mathcal{A}$ can sequentially retrieve the worth of T many, possibly duplicate, coalitions to construct its estimate $\hat{\phi}$. This captures the limitation in time, computational resources, or monetary units that a practical user is facing to avoid falling victim to the exact computation's complexity. Furthermore, it is motivated by the observation that the access to ν poses a common bottleneck, by performing inference of complex models or re-training on large data, instead of the negligible arithmetic operations of $\mathcal{A}$.

Since $\mathcal{A}$ potentially uses randomization, for instance by drawing samples and evaluating random coalitions, the comparison of $\hat{\phi}$ and ϕ needs to incorporate this randomness to judge the approximation quality. In light of this, the expected *mean squared error* is a wide-spread measure of approximation quality that is to be minimized by $\mathcal{A}$:

$$\mathbb{E}[\text{MSE}] := \frac{1}{n} \sum_{i \in \mathcal{N}} \mathbb{E}\left[\left(\phi_i - \hat{\phi}_i\right)^2\right] . \qquad (2)$$

3.3. Identifying Top-k Players: A Subtle but Significant Difference

Instead of estimating the exact Shapley values of *all* players, of which many might be similar and insignificant, one could be interested in just finding the players that possess the highest Shapley values, with the particular values being incidental. More precisely, in the *top-k identification problem* (TkIP) an approximation algorithm $\mathcal{A}$ is confronted with the task of returning an estimate $\hat{\mathcal{K}} \subseteq \mathcal{N}$ of the coalition $\mathcal{K}^*$ with given size $k \in [n] := \{1, \dots, n\}$ that contains the players with the highest Shapley values in the game $(\mathcal{N}, \nu)$. We consider again the fixed-budget setting with budget T.

However, $\mathcal{K}^*$ is not necessarily unique as players may share the same Shapley value. We restrain from any assumptions

on the value function ν and will thus present notions and measures capable of handling the ambiguity of $\mathcal{K}^*$. We call a coalition $\mathcal{K} \subseteq \mathcal{N}$ of k many players *eligible* if the sum of Shapley values associated to the players in $\mathcal{K}$ is maximal:

$$\sum_{i \in \mathcal{K}} \phi_i = \max_{S \subseteq \mathcal{N} : |S| = k} \sum_{i \in S} \phi_i \,. \tag{3}$$

We denote by $\mathcal{E}_k \subseteq \mathcal{P}(\mathcal{N})$ the set of all eligible coalitions. Any eligible estimate $\hat{\mathcal{K}}$ is correct and $\mathcal{A}$ should not be punished for it. Note that for distinct Shapley values we have $\mathcal{E}_k = \{\mathcal{K}^*\}$. In the following, we give in a first step precision measures (to be maximized) and error measures (to be minimized) for $\hat{\mathcal{K}}$ given $\mathcal{E}_k$ and extend them in a second step to the randomness of $\mathcal{A}$. A straightforward way to judge the quality of an estimate $\mathcal{K}$ is the *binary precision* (Kolpaczki et al., 2021)

$$\psi_{\text{bin}}(\hat{\mathcal{K}}) := \begin{cases} 1 & \text{if } \hat{\mathcal{K}} \in \mathcal{E}_k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

that maximally punishes every wrongly included player in $\hat{\mathcal{K}}$. In order to further differentiate estimates that are close to being eligible from ones that have little overlap with an eligible coalition, we introduce the *ratio precision*

$$\psi_{\text{rat}}(\hat{\mathcal{K}}) := \frac{1}{k} \max_{\mathcal{K} \in \mathcal{E}_k} |\mathcal{K} \cap \hat{\mathcal{K}}| \tag{5}$$

which measures the percentage of correctly identified players in $\hat{\mathcal{K}}$ by counting how many players can remain in $\hat{\mathcal{K}}$ after swapping with players from $\mathcal{N} \setminus \hat{\mathcal{K}}$ to form an eligible coalition. It serves as a gradual but still discrete refinement of the binary precision with both measures assigning values in the unit interval $[0, 1]$. Let $\phi_{k^*} := \min_{\mathcal{K} \in \mathcal{E}_k} \min_{i \in \mathcal{K}} \phi_i$ be the minimal Shapley value in any eligible coalition. Obviously, it is the minimal value for all coalitions in $\mathcal{E}_k$. Kariyappa et al. (2024) propose the *inclusion-exclusion error* which is the smallest $\varepsilon > 0$ that fulfills

$$\underbrace{\phi_i \geq \phi_{k^*} - \varepsilon}_{\text{inclusion}} \quad \text{and} \quad \underbrace{\phi_j \leq \phi_{k^*} + \varepsilon}_{\text{exclusion}} \tag{6}$$

for all $i \in \hat{\mathcal{K}}$ and all $j \in \mathcal{N} \setminus \hat{\mathcal{K}}$:

$$\rho_{\text{inc+exc}} := \inf\{\varepsilon \in \mathbb{R}^{\geq 0} \mid \forall i \in \hat{\mathcal{K}} : \phi_i \geq \phi_{k^*} - \varepsilon, \\ \forall j \in \mathcal{N} \setminus \hat{\mathcal{K}} : \phi_j \leq \phi_{k^*} + \varepsilon\} \,. \tag{7}$$

In simple terms, it measures how much the sum of Shapley values associated with $\hat{\mathcal{K}}$ can increase at least or that of $\mathcal{N} \setminus \hat{\mathcal{K}}$ can decrease by swapping a single player between them. To account for the randomness of $\mathcal{A}$, effectively turning $\hat{\mathcal{K}}$ into a random variable, the expectation of each measure poses a reasonable option just as in Section 3.2. Worth mentioning is that $\mathbb{E}[\psi_{\text{bin}}(\hat{\mathcal{K}})]$ turns out to be the probability that $\mathcal{A}$ flawlessly solves the top-k identification problem. Kariyappa et al. (2024) resort to probably approximate correct (PAC) learning. Specifically for the inclusion-exclusion error they call $\mathcal{A}$ for $\delta \in [0, 1]$ an (ϵ, δ)-PAC learner if

$$\mathbb{P}(\rho_{\text{inc+exc}}(\hat{\mathcal{K}}) \leq \varepsilon) \geq 1 - \delta \tag{8}$$

holds after $\mathcal{A}$ terminates on its own with unlimited budget at disposal. Obviously, any algorithm for the approximate-all problem can be translated to top-k identification by simply returning the k players with the highest estimates.

4. The Opportunity of Correlated Observations

The two problems of approximating all players and top-k identification differ in goal and quality measures, hence they also incentivize different sampling schemes. It is the aim of our work to emphasize and draw attention to our observation that the role of correlated samples between players plays a fundamental role for the top-k identification problem, whereas this is not the case for the approximate-all problem. We demonstrate this at the example of a simple and special class of approximation algorithms that can solve both problem statements. We call an algorithm $\mathcal{A}$ an *unbiased equifrequent player-wise independent sampler* if it samples marginal contributions for all players in M many rounds. In each round $m \in \{1, \ldots, M\}$ $\mathcal{A}$ draws n coalitions $S_1^{(m)}, \ldots, S_n^{(m)}$, one for each $i \in \mathcal{N}$, according to a fixed joint probability distribution over $\mathcal{P}(\mathcal{N} \setminus \{1\}) \times \ldots \times \mathcal{P}(\mathcal{N} \setminus \{n\})$ with marginal distribution

$$\mathbb{P}\left(S_i^{(m)} = S\right) = \frac{1}{n \cdot \binom{n-1}{|S|}} \tag{9}$$

for each $i \in \mathcal{N}$. Note that this implies $\mathbb{E}[\Delta_i(S_i^{(m)})] = \phi_i$ for all players. Further, the samples are independent between rounds and $\mathcal{A}$ aggregates the samples of each player to an estimate of its Shapley value $\hat{\phi}_i$ by taking the mean of their resulting marginal contributions, i.e.

$$\hat{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} \Delta_i\left(S_i^{(m)}\right) \,, \tag{10}$$

which is an unbiased estimate of ϕ_i. For the approximate-all problem $\mathcal{A}$ simply returns these estimates and for identifying the top-k players it returns the set of k players $\hat{\mathcal{K}}$ that yield the highest estimates $\hat{\phi}_i$. Ties can be solved arbitrarily. A well-known member of this class of approximation algorithms is *ApproShapley* proposed by Castro et al. (2009). For the approximate-all problem one can quickly derive the expected mean squared error of $\mathcal{A}$ to be

$$\mathbb{E}[\text{MSE}] = \frac{1}{nM} \sum_{i \in \mathcal{N}} \sigma_i^2 \,, \tag{11}$$

where $\sigma_i^2 := \mathbb{V}[\Delta_i(S_i^{(m)})]$ denotes the variance of player i's marginal contributions. The expected MSE decreases for a growing number of samples M and the sum of variances σ_i^2 can be seen as a constant property of the game $(\mathcal{N}, \nu)$ that is independent of $\mathcal{A}$. In contrast, turning to top-k identification, we show the emergence of another quantity in Theorem 4.1 if one considers the inclusion-exclusion error. Let $\mathbb{K}_\varepsilon := \{\mathcal{K} \subseteq \mathcal{N} \mid |\mathcal{K}| = k, \rho_{\text{inc+exc}}(\mathcal{K}) \leq \varepsilon\}$ for any $\varepsilon \in \mathbb{R}^{\geq 0}$. The central limit theorem can be applied within our considered class and thus we assume each $\sqrt{M}((\hat{\phi}_i - \hat{\phi}_j) - (\phi_i - \phi_j))$ to be normally distributed.

Theorem 4.1. *Every unbiased equifrequent player-wise independent sampler $\mathcal{A}$ for the top-k identification problem returns for any cooperative game $(\mathcal{N}, \nu)$ an estimate $\hat{\mathcal{K}}$ with inclusion-exclusion error of at most $\varepsilon \geq 0$ with probability at least*

$$\mathbb{P}(\hat{\mathcal{K}} \in \mathbb{K}_\varepsilon) \geq \sum_{\mathcal{K} \in \mathbb{K}_\epsilon} \left[1 - \sum_{\substack{i \in \mathcal{K} \\ j \in \mathcal{N} \setminus \mathcal{K}}} \Phi\left(\sqrt{M} \frac{\phi_j - \phi_i}{\sigma_{i,j}} \right) \right],$$

where $\sigma_{i,j}^2 := \mathbb{V}[\Delta_i(S_i^{(m)}) - \Delta_j(S_j^{(m)})]$ and Φ denotes the standard normal cumulative distribution function.

The proof is given in Appendix A.1. Notice the difference to Equation (11) for approximating all Shapley values. The MSE directly reflects the change of each single player's estimate $\hat{\phi}_i$, but in contrast, for identifying top-k an estimate may change arbitrarily as long as the partitioning of the players into top-k and outside of top-k stays the same.

For most pairs i, j with $i \in \mathcal{K}$ and $j \in \mathcal{N} \setminus \mathcal{K}$ of a coalition $\mathcal{K} \in \mathbb{K}_\varepsilon$ with sufficiently small ϵ, it holds $\phi_i > \phi_j$. Thus, for a fixed game $(\mathcal{N}, \nu)$ and fixed budget T, the lower bound in Theorem 4.1 should favorably increase if $\sigma_{i,j}$ decreases which can be influenced by $\mathcal{A}$ due to the allowed flexibility in its sampling scheme. Note that $\mathcal{A}$ is only restricted in the marginal contribution of each $S_i^{(m)}$ but not in the joint distribution of $S_1^{(m)}, \ldots, S_n^{(m)}$. In fact, the variance of the difference between marginal contributions decomposes to

$$\sigma_{i,j}^2 = \sigma_i^2 + \sigma_j^2 - 2\text{Cov}\left(\Delta_i\left(S_i^{(m)}\right), \Delta_j\left(S_j^{(m)}\right)\right). \quad (12)$$

Consequently, an increased covariance between sampled marginal contributions of top-k players and bottom players improves our lower bound. Leveraging the impact of covariance shown by Equation (12) in the sampling procedure is generally known as *antithetic sampling*, a variance reduction technique for Monte Carlo methods to which our class belongs. Our considered class of approximation algorithms does not impose any restrictions on the contained covariance between marginal contributions sampled within the same round m. We interpret this as degrees of freedom to shape the sampling distribution. Striving towards more reliable estimates $\hat{\mathcal{K}}$, we propose in Section 5 an approach based on the suspected improvement that positively correlated observations promise.

5. Antithetic Sampling Approach

Motivated by Section 4, we develop in Section 5.1 *Comparable Marginal Contributions Sampling* (CMCS), a budget-efficient antithetic sampling procedure that naturally yields correlated observations applicable for both problem statements. We take inspiration from (Kolpaczki et al., 2021; Kariyappa et al., 2024) and extend CMCS with a greedy selection criterion in Section 5.2, deciding from which players to sample from, to exploit opportunities that top-k identification offers.

5.1. Sampling Comparable Marginal Contributions

We start by observing that the sampling of marginal contributions can be designed to consume less than two evaluations of ν per sample. In fact, the budget restriction T is not coupled to the evaluation of marginal contributions as atomic units but single accesses to ν. Instead of separately evaluating $\nu(S)$ and $\nu(S \cup \{i\})$ for each $\Delta_i(S)$, the evaluations can be reused to form other marginal contributions and thus save budget. This idea can already be applied to the sampling of permutations of the player set. Castro et al. (2009) evaluate for each drawn permutation π the marginal contribution $\Delta_i(\text{pre}_i(\pi))$ of each player i to the preceding players in π. Except for the last player in π, each evaluation $\nu(\text{pre}_i(\pi) \cup \{i\})$ can be reused for the marginal contribution of the succeeding player.

We further develop this paradigm of *sample reusage* by exploiting the fact that any coalition $S \subseteq \mathcal{N}$ appears in n many marginal contributions, one for each player, namely in $n - |S|$ many of the form $\Delta_i(S)$ for $i \notin S$ and $|S|$ many of the form $\Delta_i(S \setminus \{i\})$ for $i \in S$. We meaningfully unify both cases by establishing the notion of an *extended marginal contribution* in Definition 5.1.

Definition 5.1. For any cooperative game $(\mathcal{N}, \nu)$, the extended marginal contribution of a player $i \in \mathcal{N}$ to a coalition $S \subseteq \mathcal{N}$ is given by

$$\Delta_i'(S) := \nu(S \cup \{i\}) - \nu(S \setminus \{i\}).$$

Fittingly, this yields $\Delta_i'(S) = \Delta_i(S \setminus \{i\})$ for $i \in S$ and $\Delta_i'(S) = \Delta_i(S) = \Delta_i(S \setminus \{i\})$ for $i \notin S$. Thus, we circumvent the case of $\Delta_i(S) = 0$ for $i \in S$.

We aim to draw in each round m (of M many) a coalition $S^{(m)} \subseteq \mathcal{N}$, compute the extended marginal contributions $\Delta_i'(S^{(m)})$ of *all* players as illustrated in Figure 1, and update

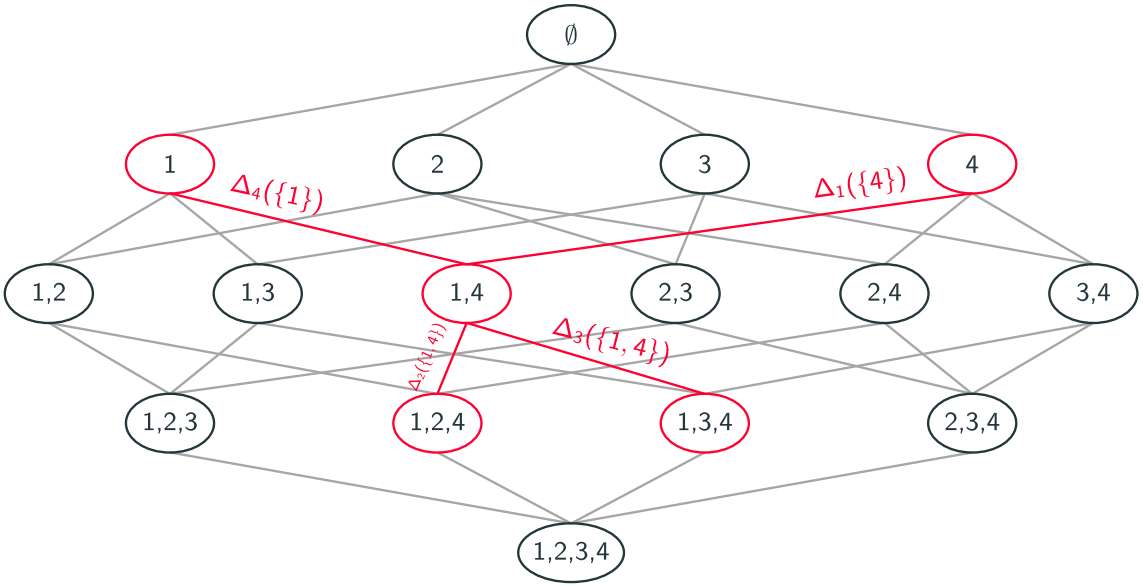


Figure 1. A cooperative game spans a lattice with each coalition $S \subseteq \mathcal{N}$ forming a node and each marginal contribution $\Delta_i(S)$ being represented by an edge between S and $S \cup \{i\}$, exemplified here for $\mathcal{N} = \{1, 2, 3, 4\}$. CMCS draws a random coalition S and computes the extended marginal contributions $\Delta'_i(S) = \Delta_i(S \setminus \{i\})$ of all players $i \in \mathcal{N}$. For $n = 4$ it evaluates five coalitions and retrieves four marginal contributions.

each $\hat{\phi}_i$ as the average of the corresponding observations:

$$\hat{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} \Delta'_i \left(S^{(m)} \right) . \tag{13}$$

We reuse the coalition value $v_{S^{(m)}} = \nu(S^{(m)})$ to update all estimates by computing each extended marginal contribution as

$$\Delta'_i \left(S^{(m)} \right) = \begin{cases} v_{S^{(m)}} - \nu(S^{(m)} \setminus \{i\}) & \text{if } i \in S \\ \nu(S^{(m)} \cup \{i\}) - v_{S^{(m)}} & \text{otherwise} \end{cases} . \tag{14}$$

Consequently, updating all n estimates requires only $n + 1$ calls to ν such that we obtain a *budget-efficiency* of $\frac{n}{n+1}$ sampled observations per call. In comparison, drawing marginal contributions separately yields a budget-efficiency of $1/2$. In order to make this approach effective, it is desirable to obtain unbiased estimates leading to the question whether there even exists a probability distribution over $\mathcal{P}(\mathcal{N})$ to sample $S^{(m)}$ from such that $\mathbb{E}[\Delta'_i(S^{(m)})] = \phi_i$ for all $i \in \mathcal{N}$. Indeed, we show its existence in Proposition 5.2 by means of a novel representation of the Shapley value based on extended marginal contributions.

Proposition 5.2. *For any cooperative game $(\mathcal{N}, \nu)$, the Shapley value of each player $i \in \mathcal{N}$ is a weighted average of its extended marginal contributions. In particular, it holds*

$$\phi_i = \sum_{S \subseteq \mathcal{N}} \frac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta'_i(S) .$$

See Appendix A.2 for a proof. The weighted average allows to view the Shapley value as the expected extended marginal contribution and thus drawing $S^{(m)}$ from the distribution

$$\mathbb{P}\left(S^{(m)} = S \right) = \frac{1}{(n+1)\binom{n}{|S|}} \quad \text{for all } S \subseteq \mathcal{N} \tag{15}$$

yields unbiased estimates. Note that this is indeed a well-defined probability distribution over $\mathcal{P}(\mathcal{N})$ as shown in Appendix A.2. The resulting algorithm *Comparable Marginal Contributions Sampling* (CMCS) is given by Algorithm 1. It requires the cooperative game (N, ν), the budget T, and the parameter k as input. The number of performed rounds M is bounded by $M = \lfloor \frac{T}{n+1} \rfloor$. We solve sampling from the exponentially large power set of $\mathcal{N}$ by first drawing a size ℓ ranging from 0 to n uniformly at random (line 3) and then drawing uniformly a coalition S of size ℓ (line 4). This results in the probability distribution of Equation (15) since there are $n + 1$ sizes and $\binom{n}{\ell}$ coalitions of size ℓ to choose from. For the top-k identification problem CMCS returns the set of k many players $\hat{\mathcal{K}}$ for which it maintains the highest estimates $\hat{\phi}_i$. Ties are solved arbitrarily.

CMCS can also be applied for the approximate-all problem by simply returning its estimates since its sampling procedure and computation of estimates is independent of k. Thus, it is also an unbiased equifrequent player-wise independent sampler (see Section 4) because the marginal contributions obtained in each round stem from a fixed joint distribution and the resulting marginal distributions coincide with Equation (9) as implied by Proposition 5.2. Hence for

Algorithm 1 Comparable Marginal Contributions Sampling (CMCS)

Input: $(\mathcal{N}, \nu)$, $T \in \mathbb{N}, k \in [n]$

$\hat{\phi}_i \leftarrow 0$ for all $i \in \mathcal{N}$
for $m = 1, \dots, \lfloor \frac{T}{n+1} \rfloor$ **do**
  Draw $\ell \in \{0, \dots, n\}$ uniformly at random
  Draw $S \subseteq \mathcal{N}$ with $|S| = \ell$ uniformly at random
  $v_S \leftarrow \nu(S)$
  for $i \in \mathcal{N}$ **do**
    $\Delta_i \leftarrow \begin{cases} v_S - \nu(S \setminus \{i\}) & \text{if } i \in S \\ \nu(S \cup \{i\}) - v_S & \text{otherwise} \end{cases}$
    $\hat{\phi}_i \leftarrow \frac{(m-1)\cdot\hat{\phi}_i + \Delta_i}{m}$
  end for
end for

Output: $\hat{\mathcal{K}}$ containing k players with highest estimate $\hat{\phi}_i$

T being a multiple of $n+1$, its expected MSE is according to Equation (11):

$$\mathbb{E}[\text{MSE}] = \frac{n+1}{nT} \sum_{i \in \mathcal{N}} \sigma_i^2 \,. \tag{16}$$

For the top-k identification the sampling scheme in CMCS yields an interesting property. All players share extended marginal contributions to the same reference coalitions $S^{(m)}$. Intuitively, this makes the estimates more comparable, as all have been updated using the same samples. Instead of estimating ϕ_i and ϕ_j precisely, CMCS answers the relevant question whether $\phi_i > \phi_j$ holds, by comparing the players marginal contributions to roughly the same coalitions, modulo the case of $i \in S$ and $j \notin S$ or vice versa. Instead, drawing marginal contributions separately, independently between the players, can, metaphorically speaking, be viewed as comparing apples with oranges.

Consequently, the estimates $\hat{\phi}_i$ and $\hat{\phi}_j$ are correlated and we further conjecture that the covariance $\text{Cov}(\Delta_i'(S^{(m)}), \Delta_j'(S^{(m)})) = \mathbb{E}[\Delta_i'(S^{(m)})\Delta_j'(S^{(m)})] - \mathbb{E}[\Delta_i'(S^{(m)})]\mathbb{E}[\Delta_j'(S^{(m)})]$ has a positive impact on the inclusion-exclusion error of CMCS in light of Theorem 4.1. For cooperative games in which the marginal contribution of a player is influenced by the coalitions size, our sampling scheme should yield positively correlated samples. In this case, if player i or j is added to the same coalition S, it is likely that both have a positive marginal contribution (or both share a negative) which in turn speaks for a positive covariance. For the general case, the covariance is stated in Proposition 5.3.

Proposition 5.3. *For any cooperative game $(\mathcal{N}, \nu)$ the covariance between the extended marginal contributions of any players $i \neq j$ of the same round sampled by CMCS is given by*

$$\text{Cov}\left(\Delta_i'\left(S^{(m)}\right), \Delta_j'\left(S^{(m)}\right)\right) = \frac{1}{n+1} \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \Delta_i(S) \left(\frac{\Delta_j'(S)}{\binom{n}{|S|}} + \frac{\Delta_j'(S \cup \{i\})}{\binom{n}{|S|+1}} \right) - \phi_i \phi_j \,.$$

The proof is given in Appendix A.2. The sum can be seen as the Shapley value ϕ_i in which each marginal contribution of i is additionally weighted by extended marginal contributions of j. To demonstrate the presumably positive covariance and give evidence to our conjecture, we consider a simple game of arbitrary size n with $\nu(\mathcal{N}) = 1$ and $\nu(S) = 0$ for all coalitions $S \neq \mathcal{N}$. Each player has a Shapley value of $\frac{1}{n}$ and the covariance in Proposition 5.3 given by $\frac{1}{n+1} - \frac{1}{n^2}$ is strictly positive for $n \geq 2$.

5.2. Relaxed Greedy Player Selection for Top-k Identification

Striving for budget-efficiency in the design of a sample procedure might be favorable, however, CMCS as proposed in Section 5.1 is forced to spend budget on the retrieval of marginal contributions for all players in order to maximize budget-efficiency. This comes with the disadvantage that evaluations of ν are performed to sample for a player i whose estimate $\hat{\phi}_i$ is possibly already reliable enough and does not need further updates compared to other players. This does not even require $\hat{\phi}_i$ to be precise in absolute terms. Instead, it is sufficient to predict with certainty whether i belongs to the top-k or not by comparing it to the other estimates. This observation calls for a more selective mechanism deciding which players to leave out in each round and thus save budget.

A radical approach is the greedy selection of a single player which maximizes a *selection criterion* based on the collected observations that incorporates incentives for exploration and exploitation. Gap-E (Gabillon et al., 2011; Bubeck et al., 2013) composes the selection criterion out of the uncertainty of a player's top-k (exploitation) membership and its number of observations (exploration). Similarly, BUS (Kolpaczki et al., 2021) selects the player i minimizing the product of its estimate's distance to the predicted top-k border $\frac{1}{2}(\min_{i \in \hat{\mathcal{K}}} \hat{\phi}_i - \max_{j \in \mathcal{N} \setminus \hat{\mathcal{K}}} \hat{\phi}_j)$ times its sample number M_i. In the same spirit but outside of the fixed-budget setting, SHAP@K (Kariyappa et al., 2024) chooses for given $\delta \in (0, 1)$ the two players $i \in \hat{\mathcal{K}}$ and $j \in \mathcal{N} \setminus \hat{\mathcal{K}}$ with the highest overlap in their δ/n-confidence intervals of their estimates $\hat{\phi}_i$ and $\hat{\phi}_j$. It applies a stopping condition and terminates when no overlaps between $\hat{\mathcal{K}}$ and $\mathcal{N} \setminus \hat{\mathcal{K}}$ larger then a specified error ε exist. Assuming normally distributed estimates $\hat{\phi}_i$ under the central limit theorem, it holds $\mathbb{P}(\rho_{\text{inc+exc}}(\hat{\mathcal{K}}) \leq \varepsilon) \geq 1 - \delta$ for its prediction $\hat{\mathcal{K}}$.

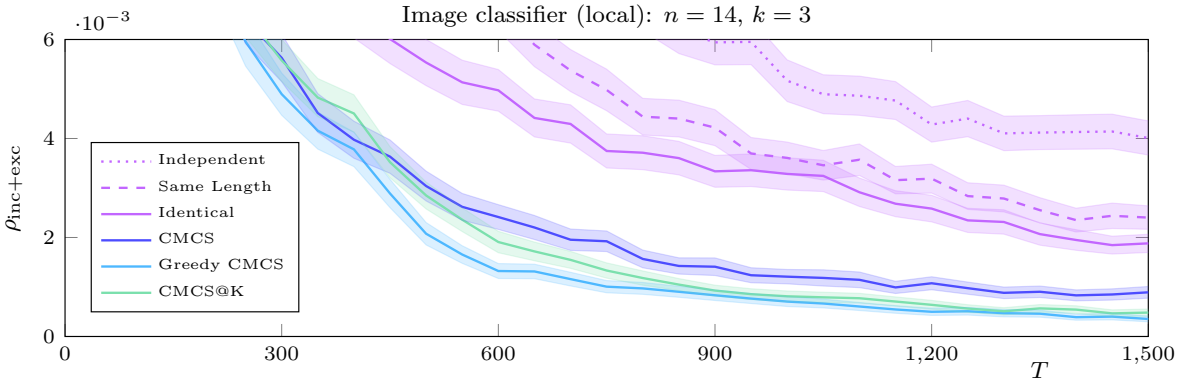


Figure 2. Inclusion-exclusion error ε for increasingly comparable sampling variants (*Independent*, *Same Length*, *Identical*), incorporation of sample-reusage (CMCS), and greedy selection (Greedy CMCS, CMCS@K) depending on T.

Given the core idea of CMCS to draw samples for multiple players at once in order to increase budget-efficiency and obtain correlated observations, the greedy selection of a single player as done in (Gabillon et al., 2011; Kolpaczki et al., 2021) or just a pair (Kariyappa et al., 2024) is not suitable for our method. The phase-wise elimination performed by SAR (Bubeck et al., 2013) is not viable as it assumes all observations to be independent in order to analytically derive phase lengths. Instead, we relax the greediness by probabilistically selecting a set of players $P^{(m)} \subseteq \mathcal{N}$ in each round m, favoring those players who fulfill a selection criterion to higher degree. By doing so, we propose *Greedy CMCS* that intertwines the overcoming of the exploration-exploitation dilemma with the pursuit of budget-efficiency. We do not abandon exploration, since every player gets a chance to be picked, and the selection criterion incentivizes exploitation as it reflects how much the choice of a player benefits the prediction $\hat{\mathcal{K}}$.

Our selection criterion is based on the current knowledge of $\hat{\phi}_1, \ldots, \hat{\phi}_n$ and the presumably best players $\hat{\mathcal{K}}$. Inspired by Theorem 4.1, we approximate the probability of each pair of players $i \in \hat{\mathcal{K}}$ and $j \in \mathcal{N} \setminus \hat{\mathcal{K}}$ being incorrectly partitioned by Greedy CMCS as

$$\hat{p}_{i,j} := \Phi\left(\sqrt{M_{i,j}}\frac{\hat{\delta}_{i,j}}{\hat{\sigma}_{i,j}}\right) . \tag{17}$$

For all pairs $(i,j) \in \mathcal{N}^2$ we track:

- the number of times $M_{i,j}$ that both i and j have been selected in a round,
- the sampled marginal contributions' mean difference $\hat{\delta}_{i,j} := \frac{1}{M_{i,j}} \sum_{m=1}^{M_{i,j}} \Delta'_j(S^{(f_{i,j}(m))}) - \Delta'_i(S^{(f_{i,j}(m))})$ within these $M_{i,j}$ rounds, where $f_{i,j}(m)$ denotes the m-th round in which i and j are selected, and
- the estimate $\hat{\sigma}^2_{i,j}$ of the variance $\sigma^2_{i,j} := \mathbb{V}[\Delta'_i(S^{(m)}) - \Delta'_j(S^{(m)})]$ w.r.t. Equation (15).

Important to note is that we may not simply use the difference $\hat{\phi}_j - \hat{\phi}_i$ of our Shapley estimates, including all rounds, instead of $\hat{\delta}_{i,j}$ because $\hat{\phi}_i$ and $\hat{\phi}_j$ may differ in their respective total amount of total samples M_i and M_j such that the central limit theorem used for Theorem 4.1 is not applicable anymore. We derive Equation (17) in Appendix A.3.

For each pair $(i,j) \in \hat{\mathcal{K}} \times (\mathcal{N} \setminus \hat{\mathcal{K}})$ the estimate $\hat{p}_{i,j}$ quantifies how likely i and j are wrongly partitioned: Greedy CMCS estimates $\hat{\phi}_i \geq \hat{\phi}_j$ although $\phi_i < \phi_j$ holds. Since we want to minimize the probability of such a mistake, it comes natural to include the pair (i,j) with the highest estimate $\hat{p}_{i,j}$ in the next round of Greedy CMCS to draw marginal contributions from, i.e. $i,j \in P^{(m)}$. As a consequence, $\hat{\phi}_i$ and $\hat{\phi}_j$ should become more reliable causing the error probability to shrink. Let $Q^{(m)} \subseteq \hat{\mathcal{K}} \times (\mathcal{N} \setminus \hat{\mathcal{K}})$ be the set of selected pairs in round m from which the selected players are formed as $P^{(m)} = \{i \in \mathcal{N} \mid \exists (i,j) \in Q^{(m)} \vee \exists (j,i) \in Q^{(m)}\}$. In order to allow for more than two updated players in a round m, i.e. $|Q^{(m)}| > 1$, but waive pairs that are more likely to be correctly classified, we probabilistically include pairs in $Q^{(m)}$ depending on their $\hat{p}$-value. Let $\hat{p}_{\max} = \max_{i \in \hat{\mathcal{K}}, j \notin \hat{\mathcal{K}}} \hat{p}_{i,j}$ be the currently highest and $\hat{p}_{\min} = \min_{i \in \hat{\mathcal{K}}, j \notin \hat{\mathcal{K}}} \hat{p}_{i,j}$ the currently lowest value. We select each pair (i,j) independently with probability

$$\mathbb{P}\left((i,j) \in Q^{(m)}\right) = \frac{\hat{p}_{i,j} - \hat{p}_{\min}}{\hat{p}_{\max} - \hat{p}_{\min}} \quad \text{for all } (i,j) \in \hat{\mathcal{K}} \times (\mathcal{N} \setminus \hat{\mathcal{K}}) . \tag{18}$$

This forces the pair with $\hat{p}_{\max}$ to be picked and that with $\hat{p}_{\min}$ to be left out. The probability of a pair beings elected increases monotonically with its $\hat{p}$-value.

Within an executed round we do not only collect marginal contributions for players in $P^{(m)}$ and update $M_{i,j}$, $\hat{\delta}_{i,j}$, and $\hat{\sigma}^2_{i,j}$ for all $(i,j) \in Q^{(m)}$. We use the collected information to its fullest by also updating the estimates of all

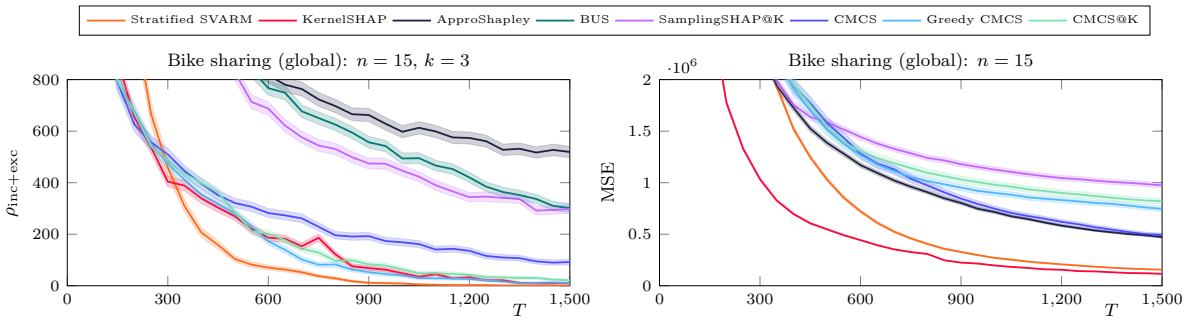


Figure 3. Comparison of achieved inclusion-exclusion error of various algorithms for top-k identification (left) and approximate-all (right) depending on T.

pairs (i, j) with both players being present in $P^{(m)}$ despite $(i, j) \notin Q^{(m)}$. Visually speaking, we update the complete subgraph induced by $P^{(m)}$ with players being nodes and edges containing the pairwise estimates.

Since the assumption of normally distributed estimates motivated by the central limit theorem is not appropriate for a low number of samples, we initialize Greedy CMCS with a warm-up phase as proposed for SHAP@K (Kariyappa et al., 2024). During the warm-up M_{min} many rounds of CMCS are performed such that afterwards every player's Shapley estimate is based on M_{min} samples. This consumes a budget of $(n + 1)M_{\text{min}}$ many evaluations. M_{min} is provided to Greedy CMCS as a parameter. Subsequently, the above described round-wise greedy sampling is applied as the second phase until the depletion of the in total available budget T. The pseudocode of the resulting algorithm Greedy CMCS is given in Appendix B.

Instead of our proposed selection mechanism, one can sample in the second phase only from the two players $i \in \hat{\mathcal{K}}$ and $j \notin \hat{\mathcal{K}}$ with the biggest overlap in confidence intervals as performed by SHAP@K. Leaving the sampling of CMCS in the first phase untouched, we call this variant *CMCS@K*. This is feasible since the choice of the sampling procedure in SHAP@K is to some extent arbitrary, as long as it yields confidence intervals for the Shapley estimates.

6. Empirical Results

We conduct multiple experiments of different designs to assess the performance of sampling comparable marginal contributions at the example of explanation tasks on real-world datasets. First, we demonstrate in Section 6.1 the iterative improvements of our proposed algorithmic tricks ranging from the naive independent sampling to Greedy CMCS and CMCS@K. Section 6.2 investigates whether favorable MSE values of algorithms for the approximate-all problem translate on the same cooperative games to the inclusion-exclusion error for top-k identification. In Section 6.3 we compare our variants of CMCS against baselines and state-of-the-art competitors. Lastly, we investigate in Section 6.4 the required budget until the stopping criterion of (Kariyappa et al., 2024) applied to CMCS guarantees an error of at most ε with probability at least $1 - \delta$. All performance measures are calculated by exhaustively computing the Shapley values in advance and averaging the results over 1000 runs. Standard errors are included as shaded bands. We compare against *ApproShapley* (Castro et al., 2009), *KernelSHAP* (Lundberg & Lee, 2017) (with reference implementation provided by the `shap` python package, the one to sample without replacement), *Stratified SVARM* (Kolpaczki et al., 2024a), *BUS* (Kolpaczki et al., 2021), and *SamplingSHAP@K* (Kariyappa et al., 2024) which is SHAP@K drawing samples according to ApproShapley. For both SamplingSHAP@K and CMCS@K, we use $M_{\text{min}} = 30$ and confidence intervals of δ/n with $\delta = 0.001$. We drop *Gap-E* (Gabillon et al., 2011) and *SAR* (Bubeck et al., 2013) due to worse performances [1].

Datasets and games. Analogously to (Kolpaczki et al., 2024a;b), we generate cooperative games from two types of explanation tasks in which the Shapley values represent feature importance scores. For global games, we construct the value function by training a *sklearn* random forest with 20 trees on each feature subset and taking its classification accuracy, or the R^2-metric for regression tasks, on a test set as the coalitions' worth. We employ the *Adult* ($n = 14$, classification), *Bank Marketing* ($n = 16$, classification), *Bike Sharing* ($n = 15$, regression), *Diabetes* ($n = 10$, regression), *German Credit* ($n = 20$, classification), *Titanic* ($n = 11$, classification), and *Wine* ($n = 13$, classification) dataset. For local games, we create a game by picking a random datapoint and taking a pretrained model's prediction value as each coalition's worth. Feature values are imputed by their mean, respectively mode. For this purpose we take the *Adult* ($n = 14$, XGBoost, classification), *ImageNet* ($n = 14$, ResNet18, classification), and *NLP Sentiment* ($n = 14$, DistilBERT transformer, regression, IMDB data) dataset.

[1] All code can be found at `https://github.com/timnielen/top-k-shapley`

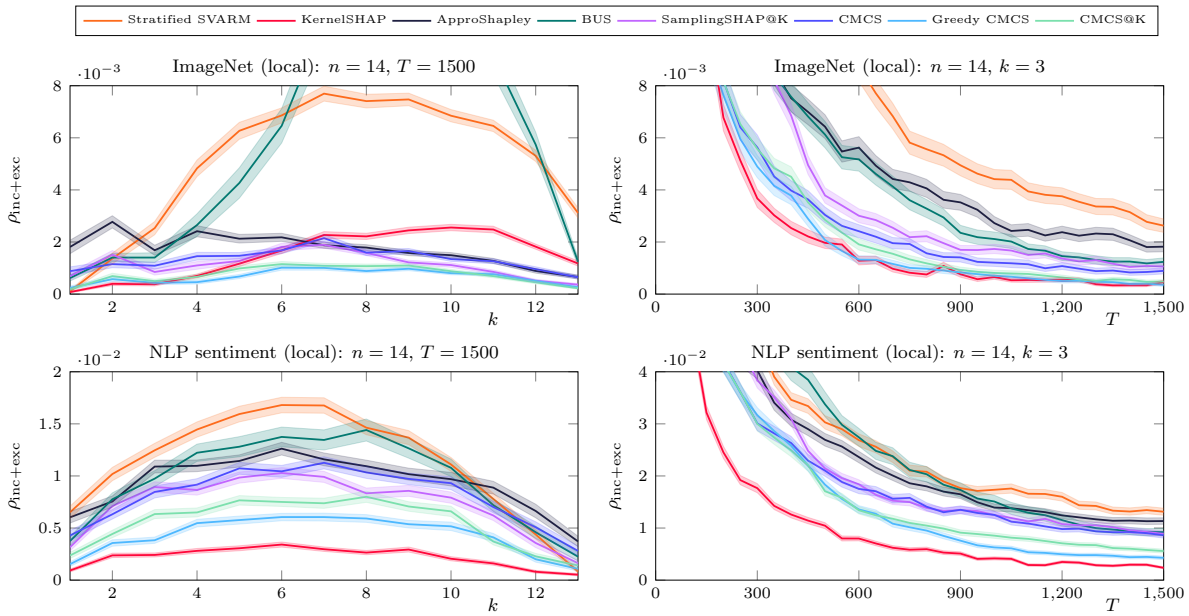


Figure 4. Comparison of achieved inclusion-exclusion error with baselines for local explanations: fixed budget with varying k (left) and fixed k with increasing budget (right).

6.1. Advantage of Comparable Sampling

Greedy CMCS builds upon multiple ideas whose effects onto the approximation quality is depicted in isolation by Figure 2. As a baseline we consider the *independent* sampling of marginal contributions of each player with distribution given in Equation (9). The comparability of the samples is stepwise increased by sampling in each round marginal contributions to coalitions of the *same length* for all players, and next using the *identical* coalition $S^{(m)}$ drawn according to Equation (15). In compliance with our conjecture, the decreasing error from *independent* to *same length* and further to *identical* speaks in favor of the beneficial impact that comes with correlated observations. The biggest leap in performance is caused by reusing the evaluated worth $\nu(S^{(m)})$ appearing in each marginal contribution of the *independent* variant resulting in CMCS. The sample reusage alone almost doubles the budget-efficiency from $1/2$ to $n/n+1$. On top of that, incorporating (relaxed) greedy sampling gifts Greedy CMCS and CMCS@K a further advantage by halving the error for higher budget ranges.

6.2. MSE vs. Inclusion-Exclusion Error

Given the similarities between the problem statements of approximating all Shapley values (cf. Section 3.2) and that of top-k identification (cf. Section 3.3) at first sight, one might suspect that approximation algorithms performing well in the former, also do so in the latter and vice versa. However, Figure 3 shows a different picture. The best performing methods Stratified SVARM and KernelSHAP remain consistent but change in order. The variants of CMCS are less favorable in terms of MSE but are barely outperformed in top-k identification. We interpret this as further evidence that top-k identification indeed rewards positively correlated samples supporting our intuition of comparability. Most striking is the difference between ApproShapley and CMCS. Assuming to know $\nu(\emptyset) = 0$, ApproShapley exhibits a budget-efficiency of 1 as it consumes in each sampled permutation n evaluations and retrieves n marginal contributions, which is only slightly better than that of CMCS with $n/n+1$. Thus, it should be only marginally better in approximation according to Equation (11) and Equation (16). Our results in Figure 3 confirm the precision of our theory. However, notice how CMCS significantly outperforms ApproShapley in terms of $\rho_{\text{inc+exc}}$ despite the almost identical budget usage. Hence, it is the stronger correlation of samples drawn by CMCS combined with the nature of top-k identification that causes the observed advantage of comparable sampling.

6.3. Comparison with Existing Methods

Figure 4 and 5 compare the performances of our methods against baselines for local and global games. For fixed $k = 3$, we observe the competitiveness of Greedy CMCS and CMCS@K being mostly on par with KernelSHAP, but getting beaten by Stratified SVARM for global games, which in turns subsides at local games. Greedy CMCS exhibits stable performance across both explanation types and the whole range of k. On the other hand, if instead the budget is fixed, Greedy CMCS has often the upper hand for values of k close to $n/2$ and is even with KernelSHAP for lower k.

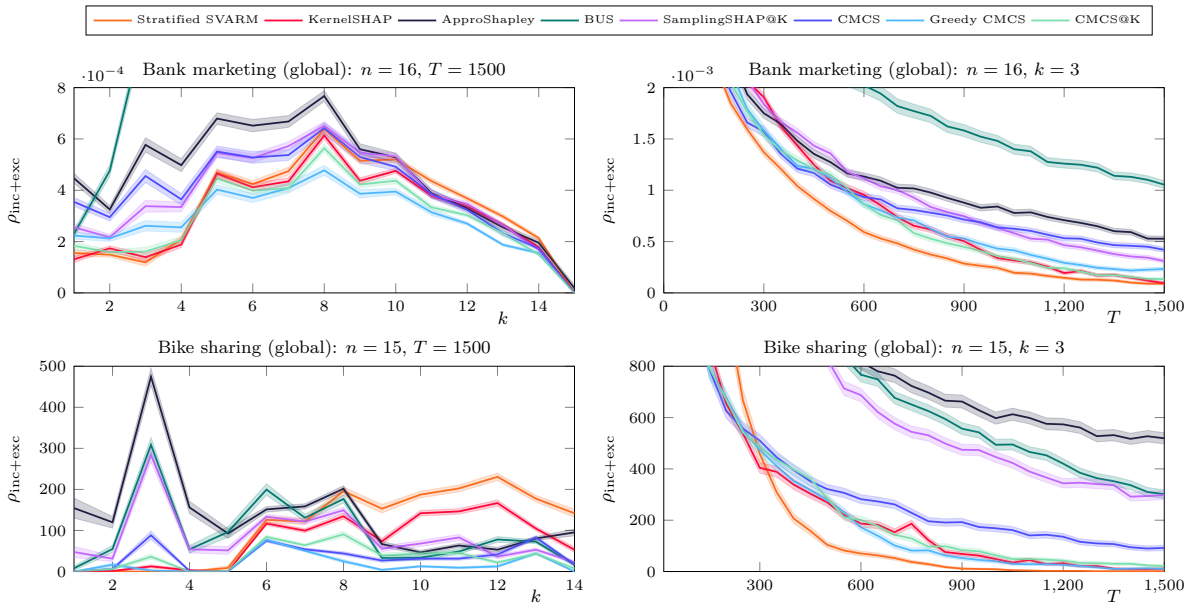


Figure 5. Comparison of achieved inclusion-exclusion error with baselines for global explanations: fixed budget with varying k (left) and fixed k with increasing budget (right).

6.4. Budget Consumption for PAC Solution

Assuming normally distributed Shapley estimates, SHAP@K is a (ε, δ)-PAC learner (Kariyappa et al., 2024), i.e. upon self-induced termination it holds $\rho_{\text{inc+exc}}(\hat{\mathcal{K}}) \leq \varepsilon$ with probability at least $1-\delta$. KernelSHAP is not applicable as it does not yield confidence bounds. For this reason Kariyappa et al. (2024) sample marginal contributions referred as *SamplingSHAP@K*. Its stopping condition is triggered as soon as no δ/n confidence intervals for the estimates $\hat{\phi}_i$ overlap between $\hat{\mathcal{K}}$ and $\mathcal{N} \setminus \hat{\mathcal{K}}$. We apply the stopping condition to our algorithms and compare to SamplingSHAP@K in the PAC-setting. Table 1 shows the average number of calls to ν until termination that is to be minimized. For some local games the number of calls is significantly higher due to the large variance in the difficulty of the respective games induced from each datapoint. CMCS@K shows the best results in nearly every game by some margin, which makes it the algorithm of choice for PAC-learning. Thus, CMCS@K is preferable when guarantees for approximation quality are required and improves upon SHAP@K due to its refined sampling mechanism.

7. Conclusion

We emphasized differences between the problem of approximating all Shapley values and that of identifying the k players with highest Shapley values. Analytically recognizing the advantage that correlated samples promise, we developed with CMCS an antithetic sampling algorithm that reuses evaluations to save budget. Our extensions Greedy CMCS and CMCS@K employ selective strategies for sampling. Both demonstrate competitive performances, with Greedy CMCS being better suited for fixed budgets, whereas CMCS@K is clearly favorable in the PAC-setting. Our proposed methods are not only model-agnostic, moreover, they can handle any cooperative game, facilitating their application for any explanation type and domain even outside of explainable AI. The difficulties that some algorithms face when translating their performance to top-k identification suggest that practitioner's being consciously interested in top-k explanations might have an advantage by applying tailored top-k algorithms instead of the trivial reduction to the approximate-all problem. Future work could investigate the sensible choice of the warm-up length in Greedy CMCS and CMCS@K which poses a trade-off between exploration and exploitation. Modifying our considered problem statement to identify the players with highest absolute Shapley values poses an intriguing variation for detecting the most impactful players and opens the door to new approaches. Finally, Shapley interactions enrich Shapley-based explanations. The number of pairwise interactions grows quadratically with n, hence top-k identification could play an even more significant role. Our work can be understood as a methodological precursor to such extensions.

References

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 258–265, 2013.

		SamplingSHAP@K		CMCS		CMCS@K		Greedy CMCS	
Game	n	#samples	SE	#samples	SE	#samples	SE	#samples	SE
Adult (global)	14	38 998	1 247	137 861	2 517	**30 995**	673	39 071	738
German credit (global)	20	21 939	336	56 738	1 129	**16 437**	248	22 327	328
Bike sharing (global)	15	4 850	97	13 053	164	**3 982**	54	8 894	117
Bank marketing (global)	16	15 124	287	39 144	875	**12 000**	206	16 260	267
Diabetes (global)	10	3 723	94	7 793	143	**2 976**	55	4 593	85
Titanic (global)	11	4 852	113	11 036	237	**3 884**	72	5 782	124
Wine (global)	13	34 953	1 046	120 859	1 906	**29 913**	641	34 265	501
NLP sentiment (local)	14	626 346	188 125	3 351 274	764 663	568 261	156 674	**447 252**	77 149
ImageNet (local)	14	135 851	39 335	578 670	196 181	**108 267**	32 067	261 586	147 126
Adult (local)	14	18 464	4 391	55 779	17 954	**14 406**	3 645	16 160	3 765

Table 1. Average number of calls to ν in the PAC-setting (see Equation (8)) across different datasets averaged over 200 runs using $\delta = 0.01$ and $\epsilon = 0.0005$ for $k = 5$.

Burgess, M. A. and Chapman, A. C. Approximating the shapley value using stratified empirical bernstein sampling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pp. 73–81, 2021.

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

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

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

Cohen, S. B., Dror, G., and Ruppin, E. Feature selection via coalitional game theory. *Neural Comput.*, 19(7):1939–1961, 2007.

Covert, I. and Lee, S.-I. Improving kernelshap: Practical shapley value estimation using linear regression. In *The 24th International Conference on Artificial Intelligence and Statistics AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3457–3465, 2021.

Covert, I., Lundberg, S., and Lee, S.-I. Shapley feature utility. In *Machine Learning in Computational Biology*, 2019.

Covert, I., Lundberg, S. M., and Lee, S. Understanding global feature contributions with additive importance measures. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Deng, X. and Papadimitriou, C. H. On the complexity of cooperative solution concepts. *Math. Oper. Res.*, 19(2): 257–266, 1994.

Doumard, E., Aligon, J., Escriva, E., Excoffier, J., Monsarrat, P., and Soulé-Dupuy, C. A comparative study of additive local explanation methods based on feature influences. In *Proceedings of the International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)*, pp. 31–40, 2022.

Gabillon, V., Ghavamzadeh, M., Lazaric, A., and Bubeck, S. Multi-bandit best arm identification. In *Proceedings in Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2222–2230, 2011.

Ghorbani, A. and Zou, J. Y. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning ICML*, volume 97, pp. 2242–2251, 2019.

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

Goldwasser, J. and Hooker, G. Stabilizing estimates of shapley values with control variates. In *Proceedings of the Second World Conference on eXplainable Artificial Intelligence (xAI)*, pp. 416–439, 2024.

Illés, F. and Kerényi, P. Estimation of the shapley value by ergodic sampling. *CoRR*, abs/1906.05224, 2019.

Kariyappa, S., Tsepenekas, L., Lécué, F., and Magazzeni, D. Shap@k: Efficient and probably approximately correct (PAC) identification of top-k features. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13068–13075, 2024.

Kolpaczki, P., Bengs, V., and Hüllermeier, E. Identifying top-k players in cooperative games via shapley bandits. In *Proceedings of the LWDA 2021 Workshops: FGWM, KDML, FGWI-BIA, and FGIR*, pp. 133–144, 2021.

Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. Approximating the shapley value without marginal contributions. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13246–13255, 2024a.

Kolpaczki, P., Haselbeck, G., and Hüllermeier, E. How much can stratification improve the approximation of shapley values? In *Proceedings of the Second World Conference on eXplainable Artificial Intelligence (xAI)*, pp. 489–512, 2024b.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4768–4777, 2017.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., and Rogers, A. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *CoRR*, abs/1306.4265, 2013.

Mitchell, R., Cooper, J., Frank, E., and Holmes, G. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022.

Molnar, C. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

Narayanam, R. and Narahari, Y. Determining the top-k nodes in social networks using the shapley value. In *Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1509–1512, 2008.

O'Brien, G., Gamal, A. E., and Rajagopal, R. Shapley value estimation for compensation of participants in demand response programs. *IEEE Transactions on Smart Grid*, 6 (6):2837–2844, 2015.

Okhrati, R. and Lipani, A. A multilinear sampling algorithm to estimate shapley values. In *25th International Conference on Pattern Recognition ICPR*, pp. 7992–7999, 2020.

Pelegrina, G. D., Kolpaczki, P., and Hüllermeier, E. Shapley value approximation based on k-additive games. *CoRR*, abs/2502.04763, 2025.

Rozemberczki, B. and Sarkar, R. The shapley value of classifiers in ensemble games. In *The 30th ACM International Conference on Information and Knowledge Management CIKM*, pp. 1558–1567, 2021.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. The shapley value in machine learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence IJCAI*, pp. 5572–5579, 2022.

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

van Campen, T., Hamers, H., Husslage, B., and Lindelauf, R. 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, 2018.

Vilone, G. and Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

Wang, H., Liang, Q., Hancock, J. T., and Khoshgoftaar, T. M. Feature selection strategies: a comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, 11(1):44, 2024.

12 Conclusion and Outlook

To conclude this thesis, we recapitulate its contributions and point out limitations of our developed methods. To address the arising challenges, we outline potential solutions, leading to future research directions. We briefly touch upon methodological extensions and variations of our considered problem statements that promise to further advance the field.

Shapley Value Approximation via Stratification. Deviating from the ubiquitous notion of marginal contributions, we proposed with *Stratified SVARM* an approximation algorithm that conducts mean estimation of the Shapley value by sampling single coalition values. The integrated stratification by coalition size is leveraged to update all player's estimates simultaneously thus reaching a more efficient usage of the limited value function evaluations. The resulting reduction in approximation error compared to the sampling of marginal contributions, which does not allow sample reusage to this degree, becomes evident by inspecting the asymptotic behavior. This phenomena is reflected by our empirical findings, albeit state-of-the-art methods being advantageous for local explanations. However, our method proves to be favorable across various domains such as global explanations, data valuation, federated learning, and ensemble selection (Muschalik et al., 2024). The displayed advantage for highly structured games demonstrated on synthetic games aligns with the interpretation of our theoretical result, stating how our algorithm benefits from stratification for cooperative games with low stratum variances. Our investigation on how to optimally allocate budget to coalition sizes quantifies the gap between the naive uniform allocation and the best in hindsight, taking the stratum variances into account. We closed this gap empirically by transferring adaptive approaches, resulting in a further improvement demonstrated by *Adaptive SVARM*. Remarkably, the optimal allocation pursued by *Adaptive SVARM* leads to an approximation error only weakly affected by the number of players, whereas methods based on marginal contributions seem incapable of achieving such a dependency.

Although stratification promises preciser stratum estimates at first, its adaptive version is vulnerable to an exaggeration of its own. For large enough player numbers,

the fixed budget is not sufficient to provide adequate variance estimates for the growing number of strata during the exploration phase. As a consequence, the estimated optimal allocation is polluted by this imprecision and misguides the exploitation phase, ultimately missing the true optimal allocation. As a remedy, one could coarsen the stratification by merging adjacent strata depending on the player number and budget, or even adaptively by taking the observed variances into account. More so, we suspect that weakening the granularity in heterogeneous manner would exploit how stratum means tend to differ less for larger sizes, allowing a coarser partitioning than at the lower end. Moreover, the hyperparemeter λ used to set the ratio between exploration and exploitation is susceptible to misspecification. Taking inspiration from explore-then-commit strategies of multi-armed bandits (Lattimore and Szepesvári, 2020) poses a possible solution. Lastly, our approximate-all problem statement deliberately assumes equal evaluation costs for all coalitions to ease theoretical analyses. However, this simplification is at least debatable, as one might observe how lower-sized coalitions require more imputation effort for local feature explanations, and the other way round, coalitions of larger size are more costly to evaluate in data valuation and federated learning. Fittingly, our stratifying methods already possess mechanisms to distinguish coalitions by size and are thus adjustable to the differing evaluation costs by tweaking the sample allocation.

Shapley Value Approximation via Optimization. Following a vastly different paradigm of estimation, we fit a surrogate game whose own Shapley values are immediately elicited out of its representation via weighted regression to the game whose Shapley values are to be approximated. To this end, we proposed a k-additive surrogate game that is composed of Shapley values and Shapley interactions up to a certain order k. Our theoretical result confirms the validity of our approach, stating that k-additivity is not a rigorous assumption but rather a tool to construct surrogate games of varying plasticity while retaining the ability to exactly mimic the true Shapley values. Consequently, one can interpret the solution to our resulting k-additive optimization problem as a novel representation of the Shapley value.

On one hand, having to specify the hyperparameter k confronts a user with an adequate choice to make. On the other hand, it provides the opportunity to incorporate domain knowledge to which dimension interactions play a significant role and diminish beyond that. The derivation of a theoretical guarantee for the fixed-budget setting proves to be challenging such that it remains unclear which properties of a cooperative game impact the estimates' precision. Moreover, as the surrogate game is already specified by its own interactions, we suspect our approach

to be likewise fruitful for the estimation of interactions. In other words, the solution of the k-additive optimization problem could yield the Shapley interactions of the game of interest for the right choice of weights that is to be derived analytically in future work.

Approximation of Shapley Interactions. In the same manner that Shapley interactions generalize the Shapley value, we extended *Stratified SVARM* to *SVARM-IQ* which estimates Shapley interactions of arbitrary order. Not only is our method capable of approximating any semivalue and cardinal interaction index, but it also does not require the indices of interest to be specified before approximation. Instead, by leveraging strata as universal building blocks for all indices, it enables to postpone the specification to even after sampling when the stratum estimates are aggregated to indices according to their weightings. *SVARM-IQ* proves to be competitive against state-of-the-art methods on various domains if not favorable (Muschalik et al., 2024). In particular, its achieved reduction in approximation error for interactions on image data processed by vision transformer models speaks for itself.

Although we did not propose an adaptive version as for *Stratified SVARM*, the fine granularity of the stratification already threatens its practicability for higher interaction order k since the number of strata grows exponentially with k. Despite that, even for $k = 2$, the number of strata has cubic complexity w.r.t. the player number n. Besides the incurred space complexity, the number of made observations per stratum shrinks, harming their estimates' precision. As argued above, we see a remedy in coarser stratification. Our method is only one example to showcase how algorithms for approximating the Shapely value can be transferred to Shapley interactions. Thus, it could pave the way for further methods to be proposed. In this spirit, we conjecture that the broad class of methods for the Shapley value that samples marginal contributions can be lifted to approximate interactions. Instead of utilizing the representation of interactions as a weighted sum of discrete derivatives, or on the other extreme, breaking them down to the atomic building blocks of coalition values as done by *SVARM-IQ*, one can apply the recursive nature of the discrete derivative (see Definition 2.19) to represent it by marginal contributions. This would immediately facilitate the extension of the aforementioned class to interactions.

Top-k Shapley Players Identification. The difference between the problem of approximating all players' Shapley values and that of only identifying those with the highest values comes with multiple opportunities and implications. First, we

establish a connection to multi-armed bandits following the observation how players form arms whose reward distributions can be constructed from their marginal contributions. Next, our analytical discovery that the covariance between player's Shapley value estimates impacts the identification performance forms a contrast to its irrelevance for the approximate-all problem under MSE minimization. Seeking to take advantage of the influence of covariance, we developed *CMCS* which combines dependent observations between players as a form of antithetic sampling and budget efficiency in its sampling mechanism. Facing the exploration-exploitation dilemma, we proposed further variants that selectively choose the players to sample from which promise more relevant information gain to separate the top-k from the rest.

Taking into consideration that players with negative Shapley values might actually be at least as impactful as the top-k, but only in a harmful way, we advocate for future work to investigate an altered problem statement in which the players with highest absolute Shapley values are to be identified. Lastly, we conjecture how top-k identification has an even higher relevance for Shapley interactions. Given an interaction order of interest, the sheer number of subsets of that cardinality should sufficiently impede the approximation of all their interaction terms such that top-k identification has a significant advantage in budget consumption as many interactions can be safely predicted to lie within the top-k or on the opposite side. We deem the combination of our efforts on the approximation of Shapley interactions and top-k identification for Shapley values to be a fruitful methodological precursor to this future endeavor.

Bibliography

Aas, Kjersti, Martin Jullum, and Anders Løland (2021). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *Artificial Intelligence* 298, p. 103502 (cit. on p. 34).

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin (2012). *Learning From Data*. AMLBook (cit. on p. 30).

Adadi, Amina and Mohammed Berrada (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6, pp. 52138–52160 (cit. on p. 32).

Ancona, Marco, Cengiz Öztireli, and Markus H. Gross (2019). "Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 272–281 (cit. on p. 50).

Balestra, Chiara, Florian Huber, Andreas Mayr, and Emmanuel Müller (2022). "Unsupervised Features Ranking via Coalitional Game Theory for Categorical Data". In: *Proceedings of Big Data Analytics and Knowledge Discovery - 24th International Conference (DaWaK)*. Vol. 13428. Lecture Notes in Computer Science. Springer, pp. 97–111 (cit. on pp. 37, 53).

Banzhaf, J.F. (1965). "Weighted voting doesn't work: A mathematical analysis". In: *Rutgers Law Review* 19.2, pp. 317–343 (cit. on p. 12).

Becker, Patrick and Viktor Bengs (2023). "Shapley-Based Feature Selection for Online Algorithm Selection". In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD, Revised Selected Papers, Part I*. Vol. 2133. Communications in Computer and Information Science. Springer, pp. 313–324 (cit. on p. 37).

Bordt, Sebastian and Ulrike von Luxburg (2023). "From Shapley Values to Generalized Additive Models and back". In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 709–745 (cit. on p. 54).

Brânzei, Rodica, Dinko Dimitrov, and Stef Tijs (2008). *Models in Cooperative Game Theory*. Lecture Notes in Economics. Springer Verlag (cit. on p. 6).

Bremer, Jörg and Michael Sonnenschein (2013). "Estimating Shapley Values for Fair Profit Distribution in Power Planning Smart Grid Coalitions". In: *Proceedings of Multiagent System Technologies - 11th German Conference (MATES)*. Vol. 8076. Lecture Notes in Computer Science. Springer, pp. 208–221 (cit. on pp. 1, 29).

Breuer, Nils Ole, Andreas Sauter, Majid Mohammadi, and Erman Acar (2024). “CAGE: Causality-Aware Shapley Value for Global Explanations”. In: *Proceedings of Explainable Artificial Intelligence - Second World Conference (xAI), Part III*. Vol. 2155. Communications in Computer and Information Science. Springer, pp. 143–162 (cit. on p. 35).

Bubeck, Sébastien, Rémi Munos, and Gilles Stoltz (2009). “Pure Exploration in Multi-armed Bandits Problems”. In: *Algorithmic Learning Theory*. Springer Berlin Heidelberg, pp. 23–37 (cit. on p. 57).

Bubeck, Sébastien, Tengyao Wang, and Nitin Viswanathan (2013). “Multiple Identifications in Multi-Armed Bandits”. In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 258–265 (cit. on p. 57).

Burgess, Mark Alexander and Archie C. Chapman (2021). “Approximating the Shapley Value Using Stratified Empirical Bernstein Sampling”. In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*. ijcai.org, pp. 73–81 (cit. on p. 46).

Campen, Tjeerd van, Herbert Hamers, Bart Husslage, and Roy Lindelauf (2018). “A new approximation method for the Shapley value applied to the WTC 9/11 terrorist attack”. In: *Social Network Analysis and Mining* 8.1, 3:1–3:12 (cit. on pp. 23, 29).

Castro, Javier, Daniel Gómez, Elisenda Molina, and Juan Tejada (2017). “Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation”. In: *Computers & Operations Research* 82, pp. 180–188 (cit. on pp. 46, 53).

Castro, Javier, Daniel Gómez, and Juan Tejada (2009). “Polynomial calculation of the Shapley value based on sampling”. In: *Computers & Operations Research* 36.5, pp. 1726–1730 (cit. on pp. 43, 57).

Charnes, A., B. Golany, M. Keane, and J. Rousseau (1988). “Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations”. In: *Econometrics of Planning and Efficiency*. Springer Netherlands, pp. 123–133 (cit. on pp. 48, 54).

Chu, Carlin and David Po Kin Chan (2020). “Feature Selection Using Approximated High-Order Interaction Components of the Shapley Value for Boosted Tree Classifier”. In: *IEEE Access* 8, pp. 112742–112750 (cit. on p. 39).

Cohen, Shay B., Gideon Dror, and Eytan Ruppin (2007). “Feature Selection via Coalitional Game Theory”. In: *Neural Computation* 19.7, pp. 1939–1961 (cit. on pp. 35, 37).

Covert, Ian and Su-In Lee (2021). “Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression”. In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 3457–3465 (cit. on p. 48).

Covert, Ian, Scott Lundberg, and Su-In Lee (2019). “Shapley feature utility”. In: *Machine Learning in Computational Biology* (cit. on p. 47).

Covert, Ian, Scott M. Lundberg, and Su-In Lee (2021). “Explaining by Removing: A Unified Framework for Model Explanation”. In: *Journal of Machine Learning Research* 22, 209:1–209:90 (cit. on p. 33).

– (2020). “Understanding Global Feature Contributions With Additive Importance Measures”. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 17212–17223 (cit. on p. 35).

Deng, Xiaotie and Christos H. Papadimitriou (1994). “On the Complexity of Cooperative Solution Concepts”. In: *Mathematics of Operations Research* 19.2, pp. 257–266 (cit. on pp. 2, 22).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Vol. 1. Association for Computational Linguistics, pp. 4171–4186 (cit. on p. 38).

Dubey, Pradeep, Abraham Neyman, and Robert James Weber (1981). “Value Theory without Efficiency”. In: *Mathematics of Operations Research* 6.1, pp. 122–128 (cit. on p. 11).

Fahimullah, Muhammad, Yasir Faheem, and Naveed Ahmad (2019). “Collaboration Formation and Profit Sharing Between Software Development Firms: A Shapley Value Based Cooperative Game”. In: *IEEE Access* 7, pp. 42859–42873 (cit. on pp. 1, 29).

Fumagalli, Fabian, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer (2024). “KernelSHAP-IQ: Weighted Least Square Optimization for Shapley Interactions”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 14308–14342 (cit. on p. 49).

– (2023). “SHAP-IQ: Unified Approximation of any-order Shapley Interactions”. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36. Curran Associates, Inc., pp. 11515–11551 (cit. on pp. 2, 48).

Gabillon, Victor, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck (2011). “Multi-Bandit Best Arm Identification”. In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Vol. 24. Curran Associates, Inc., pp. 2222–2230 (cit. on p. 57).

Gaskó, Noémi, Tamás Képes, Rodica Ioana Lung, and Mihai Suciu (2023). “Identification of influential nodes with Shapley Influence Maximization Extremal Optimization algorithm”. In: *Applied Soft Computing* 146, p. 110653 (cit. on pp. 1, 29).

Ghorbani, Amirata and James Y Zou (2020). “Neuron Shapley: Discovering the Responsible Neurons”. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 5922–5932 (cit. on p. 39).

– (2019). “Data Shapley: Equitable Valuation of Data for Machine Learning”. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2242–2251 (cit. on p. 38).

Gosiewska, Alicja and Przemyslaw Biecek (2020). *Do Not Trust Additive Explanations*. arXiv: 1903.11420 [cs.LG] (cit. on p. 36).

Grabisch, Michel (1997a). “Alternative representations of discrete fuzzy measures for decision making”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 5.5, pp. 587–607 (cit. on p. 54).

– (1997b). “k-order additive discrete fuzzy measures and their representation”. In: *Fuzzy Sets and Systems* 92.2, pp. 167–189 (cit. on p. 54).

Grabisch, Michel, Henri Prade, Éric Raufaste, and Patrice Terrier (2006). “Application of the Choquet integral to subjective mental workload evaluation”. In: *IFAC Proceedings Volumes* 39 (4), pp. 135–140 (cit. on p. 55).

Grabisch, Michel and Marc Roubens (1999). “An axiomatic approach to the concept of interaction among players in cooperative games”. In: *International Journal of Game Theory* 28.4, pp. 547–565 (cit. on pp. 2, 16–20).

Heskes, Tom, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen (2020). “Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models”. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 4778–4789 (cit. on p. 35).

Illés, Ferenc and Péter Kerényi (2019). “Estimation of the Shapley value by ergodic sampling”. In: *CoRR* abs/1906.05224. arXiv: 1906.05224 (cit. on p. 47).

Jia, Ruoxi, David Dao, Boxin Wang, et al. (2019). “Towards Efficient Data Valuation Based on the Shapley Value”. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1167–1176 (cit. on p. 38).

Kariyappa, Sanjay, Leonidas Tsepenekas, Freddy Lécué, and Daniele Magazzeni (2024). “SHAP@k: Efficient and Probably Approximately Correct (PAC) Identification of Top-K Features”. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI) and 36th Conference on Innovative Applications of Artificial Intelligence (IAAI) and 14th Symposium on Educational Advances in Artificial Intelligence (EAAI)*. AAAI Press, pp. 13068–13075 (cit. on pp. 27, 28, 58).

Kimms, Alf and Igor Kozeletskyi (2016). “Shapley value-based cost allocation in the cooperative traveling salesman problem under rolling horizon planning”. In: *EURO EURO Journal on Transportation and Logistics* 5.4, pp. 371–392 (cit. on pp. 1, 29).

Kojadinovic, Ivan (2003). “Modeling interaction phenomena using fuzzy measures: on the notions of interaction and independence”. In: *Fuzzy Sets and Systems* 135.3, pp. 317–340 (cit. on pp. 16, 17).

Kolpaczki, Patrick (2024). “Comparing Shapley Value Approximation Methods for Unsupervised Feature Importance”. In: *Proceedings of DataNinja sAIOnARA 2024 Conference*. BieColl – Bielefeld eCOllections, pp. 13–15.

Kolpaczki, Patrick, Viktor Bengs, and Eyke Hüllermeier (2021). “Identifying Top-k Players in Cooperative Games via Shapley Bandits”. In: *Proceedings of the LWDA 2021 Workshops: FGWM, KDML, FGWI-BIA, and FGIR*. Vol. 2993. CEUR Workshop Proceedings. CEUR-WS.org, pp. 133–144.

Kolpaczki, Patrick, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier (2024a). "Approximating the Shapley Value without Marginal Contributions". In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI) and 36th Conference on Innovative Applications of Artificial Intelligence (IAAI) and 14th Fourteenth Symposium on Educational Advances in Artificial Intelligence (EAAI)*. AAAI Press, pp. 13246–13255.

Kolpaczki, Patrick, Georg Haselbeck, and Eyke Hüllermeier (2024b). "How Much Can Stratification Improve the Approximation of Shapley Values?" In: *Proceedings of Explainable Artificial Intelligence - Second World Conference (xAI), Part II*. Vol. 2154. Communications in Computer and Information Science. Springer, pp. 489–512.

Kolpaczki, Patrick, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier (2024c). "SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification". In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 3520–3528.

Kolpaczki, Patrick, Tim Nielen, and Eyke Hüllermeier (2025). "Antithetic Sampling for Top-k Shapley Identification". In: *CoRR* abs/2504.02019. arXiv: 2504.02019.

Kumar, Indra, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler (2021). "Shapley Residuals: Quantifying the limits of the Shapley value for explanations". In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Curran Associates, Inc., pp. 26598–26608 (cit. on p. 36).

Kumar, Indra, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler (2020). "Problems with Shapley-value-based explanations as feature importance measures". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 5491–5500 (cit. on p. 36).

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press (cit. on p. 156).

LeCun, Yann, John Denker, and Sara Solla (1989). "Optimal Brain Damage". In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Vol. 2. Morgan Kaufmann, pp. 598–605 (cit. on p. 39).

Liu, Zelei, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui (2022). "GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning". In: *ACM Transactions on Intelligent Systems and Technology* 13.4, 60:1–60:21 (cit. on p. 38).

Lundberg, Scott M., Gabriel G. Erion, Hugh Chen, et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67 (cit. on p. 32).

Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. Curran Associates, Inc., pp. 4765–4774 (cit. on pp. 34, 48, 54).

Maleki, Sasan, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers (2013). "Bounding the Estimation Error of Sampling-based Shapley Value Approximation With/Without Stratifying". In: *CoRR* abs/1306.4265. arXiv: 1306.4265 (cit. on pp. 45, 46).

Mitchell, Rory, Joshua Cooper, Eibe Frank, and Geoffrey Holmes (2022). “Sampling Permutations for Shapley Value Estimation”. In: *Journal of Machine Learning Research* 23, 43:1–43:46 (cit. on p. 47).

Moehle, Nicholas, Stephen Boyd, and Andrew Ang (2022). “Portfolio Performance Attribution via Shapley Value”. In: *Journal of Investment Management* 20.3 (cit. on pp. 1, 29).

Molnar, Christoph (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. (cit. on p. 32).

Muschalik, Maximilian, Hubert Baniecki, Fabian Fumagalli, et al. (2024). “shapiq: Shapley Interactions for Machine Learning”. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmark Track*. Vol. 37. Curran Associates, Inc., pp. 130324–130357 (cit. on pp. 51, 56, 155, 157).

Narayanam, Ramasuri and Yadati Narahari (2008). “Determining the top-k nodes in social networks using the Shapley value”. In: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Vol. 3. IFAAMAS, pp. 1509–1512 (cit. on p. 57).

Neyman, Jerzy (1934). “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection”. In: *Journal of the Royal Statistical Society* 97.4, pp. 558–625 (cit. on pp. 45, 52).

O’Brien, Gearóid, Abbas El Gamal, and Ram Rajagopal (2015). “Shapley Value Estimation for Compensation of Participants in Demand Response Programs”. In: *IEEE Trans. Smart Grid* 6.6, pp. 2837–2844 (cit. on pp. 1, 29, 46).

Okhrati, Ramin and Aldo Lipani (2020). “A Multilinear Sampling Algorithm to Estimate Shapley Values”. In: *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 7992–7999 (cit. on p. 46).

Owen, Guillermo (1972). “Multilinear Extensions of Games”. In: *Management Science* 18.5, pp. 64–79 (cit. on p. 46).

Pelegrina, Guilherme Dean, Leonardo Tomazeli Duarte, Michel Grabisch, and João Marcos Travassos Romano (2020). “The multilinear model in multicriteria decision making: The case of 2-additive capacities and contributions to parameter identification”. In: *European Journal of Operational Research* 282.3, pp. 945–956 (cit. on p. 55).

Pelegrina, Guilherme Dean, Patrick Kolpaczki, and Eyke Hüllermeier (2025). “Shapley Value Approximation Based on k-Additive Games”. In: *CoRR* abs/2502.04763. arXiv: 2502.04763.

Pfannschmidt, Karlson, Eyke Hüllermeier, Susanne Held, and Reto Neiger (2016). “Evaluating Tests in Medical Diagnosis: Combining Machine Learning with Game-Theoretical Concepts”. In: *Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems - 16th International Conference (IPMU), Part I*. Vol. 610. Communications in Computer and Information Science. Springer, pp. 450–461 (cit. on pp. 35, 37).

Ray, Debraj (2007). *A game-theoretic perspective on coalition formation*. Oxford University Press (cit. on p. 8).

Rozemberczki, Benedek and Rik Sarkar (2021). “The Shapley Value of Classifiers in Ensemble Games”. In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*. CIKM ’21. ACM, pp. 1558–1567 (cit. on p. 39).

Rozemberczki, Benedek, Lauren Watson, Péter Bayer, et al. (2022). “The Shapley Value in Machine Learning”. In: *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*. ijcai.org, pp. 5572–5579 (cit. on pp. 2, 38).

Schopka, Kristian and Herbert Kopfer (2015). “Cost Allocation for Horizontal Carrier Coalitions Based on Approximated Shapley Values”. In: *Proceedings of Selected Papers of the International Conference of the German, Austrian and Swiss Operations Research Societies (GOR, ÖGOR, SVOR/ASRO)*. Operations Research Proceedings. Springer, pp. 133–140 (cit. on pp. 1, 29).

Sebastián, Carlos and Carlos E. González-Guillén (2024). “A feature selection method based on Shapley values robust for concept shift in regression”. In: *Neural Computing and Applications* 36.23, pp. 14575–14597 (cit. on p. 37).

Shalit, Haim (2021). “The Shapley value decomposition of optimal portfolios”. In: *Annals of Finance* 17.1, pp. 1–25 (cit. on p. 29).

– (2020). “Using the Shapley value of stocks as systematic risk”. In: *Journal of Risk Finance* 21.4, pp. 459–468 (cit. on p. 29).

Shapley, L. S. (1953). “A Value for n-Person Games”. In: *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, pp. 307–318 (cit. on pp. 1, 8, 10, 13).

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). “Learning Important Features Through Propagating Activation Differences”. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3145–3153 (cit. on p. 32).

Sim, Rachael Hwee Ling, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low (2020). “Collaborative Machine Learning with Incentive-Aware Model Rewards”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 8927–8936 (cit. on p. 38).

Simon, Grah and Thouvenot Vincent (2020). “A Projected Stochastic Gradient Algorithm for Estimating Shapley Value Applied in Attribute Importance”. In: *Proceedings of Machine Learning and Knowledge Extraction - 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE*. Vol. 12279. Lecture Notes in Computer Science. Springer, pp. 97–115 (cit. on p. 48).

Strumbelj, Erik and Igor Kononenko (2010). “An Efficient Explanation of Individual Classifications using Game Theory”. In: *Journal of Machine Learning Research* 11, pp. 1–18 (cit. on p. 34).

– (2014). “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and Information Systems* 41.3, pp. 647–665 (cit. on pp. 34, 44).

Sugeno, M. (1974). “Theory of fuzzy integrals and its applications”. PhD thesis. Tokyo Institute of Technology (cit. on p. 6).

Sundararajan, Mukund, Kedar Dhamdhere, and Ashish Agarwal (2020). “The Shapley Taylor Interaction Index”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 9259–9268 (cit. on p. 49).

Sundararajan, Mukund and Amir Najmi (2020). “The Many Shapley Values for Model Explanation”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 9269–9278 (cit. on pp. 2, 33, 34).

Tsai, Che-Ping, Chih-Kuan Yeh, and Pradeep Ravikumar (2023). “Faith-shap: The faithful shapley interaction index”. In: *Journal of Machine Learning Research* 24.94, pp. 1–42 (cit. on p. 49).

Valásková, Lubica and Peter Struk (2005). “Classes of fuzzy measures and distortion”. In: *Kybernetika* 41.2, pp. 205–212 (cit. on p. 7).

Vilone, Giulia and Luca Longo (2021). “Notions of explainability and evaluation approaches for explainable artificial intelligence”. In: *Information Fusion* 76, pp. 89–106 (cit. on p. 2).

Wang, Jiachen T. and Ruoxi Jia (2023). “Data Banzhaf: A Robust Data Valuation Framework for Machine Learning”. In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 6388–6421 (cit. on pp. 47, 50).

Wu, Mengmeng, Ruoxi Jia, Changle Lin, Wei Huang, and Xiangyu Chang (2023). “Variance reduced Shapley value estimation for trustworthy data valuation”. In: *Computers & Operations Research* 159, p. 106305 (cit. on p. 38).

A

Appendix to Approximating the Shapley Value without Marginal Contributions

A List of Symbols

Table 1: List of frequently symbols used throughout the paper.

Problem setting	
$\mathcal{N}$	set of players
$\mathcal{N}_i$	set of players without i
n	number of players
ν	value function
T	budget, number of allowed evaluations of ν
ϕ_i	Shapley value of player i
$\hat{\phi}_i$	estimated Shapley value of player i
SVARM	
ϕ_i^+	positive Shapley value
ϕ_i-	negative Shapley value
$\hat{\phi}_i^+$	estimated positive Shapley value
$\hat{\phi}_i^-$	estimated negative Shapley value
P^+	sampling probability distribution over coalitions to estimate ϕ_i^+
P^-	sampling probability distribution over coalitions to estimate ϕ_i^-
$\bar{T}$	remaining budget after completion of the warm-up phase
${\sigma_i^+}^2$	variance of coalition values including player i
${\sigma_i^-}^2$	variance of coalition values excluding player i
r_i^+	range of coalition values including player i
r_i^-	range of coalition values excluding player i
$\bar{m}_i^+$	number of sampled coalitions after the warm-up phase to update $\hat{\phi}_i^+$
$\bar{m}_i^-$	number of sampled coalitions after the warm-up phase to update $\hat{\phi}_i^-$
m_i^+	total number of sampled coalitions to update $\hat{\phi}_i^+$
m_i^-	total number of sampled coalitions to update $\hat{\phi}_i^-$
Stratified SVARM	
$\phi_{i,\ell}^+$	ℓ-th positive Shapley subvalue
$\phi_{i,\ell}^-$	ℓ-th negative Shapley subvalue
$\hat{\phi}_{i,\ell}^+$	estimated ℓ-th positive Shapley subvalue
$\hat{\phi}_{i,\ell}^-$	estimated ℓ-th positive Shapley subvalue
$\tilde{P}$	sampling probability distribution over coalition sizes
$\bar{T}$	remaining budget after completion of the warm-up phase
${\sigma_{i,\ell}^+}^2$	variance of coalition values in the ℓ-th stratum including player i
${\sigma_{i,\ell}^+}^2$	variance of coalition values in the ℓ-th stratum excluding player i
$r_{i,\ell}^+$	range of coalition values in the ℓ-th stratum including player i
$r_{i,\ell}^-$	range of coalition values in the ℓ-th stratum excluding player i
$\bar{m}_{i,\ell}^+$	number of sampled coalitions after the warm-up phase to update $\hat{\phi}_{i,\ell}^+$
$\bar{m}_{i,\ell}^-$	number of sampled coalitions after the warm-up phase to update $\hat{\phi}_{i,\ell}^-$
$m_{i,\ell}^+$	total number of sampled coalitions to update $\hat{\phi}_{i,\ell}^+$
$m_{i,\ell}^-$	total number of sampled coalitions to update $\hat{\phi}_{i,\ell}^-$

B Further Pseudocode

B.1 SVARM

Algorithm 3 WARMUP

for $i \in \mathcal{N}$ **do**
  Draw A^+ and A^- i.i.d. from P^w
  $\hat{\phi}_i^+ \leftarrow \nu(A^+ \cup \{i\})$
  $\hat{\phi}_i^- \leftarrow \nu(A^-)$
end for

The warm-up of SVARM samples for each player i two coalitions A^+ and A^-, both drawn i.i.d. according to the weights w_S, i.e., the probability distribution P^w, and updates $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$, which needs a budget of $2n$ in total. This ensures that each estimate is based on at least one sample.

B.2 Stratified SVARM

Algorithm 4 UPDATE(A)

1: $v \leftarrow \nu(A)$
2: **for** $i \in A$ **do**
3: $\quad \hat{\phi}_{i,|A|-1}^+ \leftarrow \frac{c_{i,|A|-1}^+ \cdot \hat{\phi}_{i,|A|-1}^+ + v}{c_{i,|A|-1}^+ + 1}$
4: $\quad c_{i,|A|-1}^+ \leftarrow c_{i,|A|-1}^+ + 1$
5: **end for**
6: **for** $i \in \mathcal{N} \setminus A$ **do**
7: $\quad \hat{\phi}_{i,|A|}^- \leftarrow \frac{c_{i,|A|}^- \cdot \hat{\phi}_{i,|A|}^- + v}{c_{i,|A|}^- + 1}$
8: $\quad c_{i,|A|}^- \leftarrow c_{i,|A|}^- + 1$
9: **end for**

Stratified SVARM's update procedure updates exactly one Shapley subvalue of each player given a coalition A. It consumes only one budget token by storing the worth of A in the variable v (line 1). The first loop increments for all players $i \in A$ their counter $c_{i,|A|-1}^+$ by 1 and updates the $|A|-1$-th positive Shapley subvalue estimate $\hat{\phi}_{i,|A|-1}^+$ to be the average over all values of coalitions which are contained in that stratum of player i. Analogously, the second loop incerements for all players i not contained in A their counter $c_{i,|A|}^-$ by 1 and updates the $|A|$-th negative Shapley subvalue estimate $\hat{\phi}_{i,|A|}^+$ to be the average over all values of coalitions which are contained in that stratum of player i.

Algorithm 5 EXACTCALCULATION$(\mathcal{N})$

for $s \in \{1, n-1, n\}$ **do**
  for $A \in \{S \subseteq \mathcal{N} \mid |S| = s\}$ **do**
    UPDATE(A)
  end for
end for

The exact calculation evaluates all coalitions of size $1, n-1$ and n, thus $2n+1$ in total. For each coalition, the update procedure is called. Effectively, this leads to exactly computed strata $\hat{\phi}_{i,0}^+ = \phi_{i,0}^+, \hat{\phi}_{i,n-2}^+ = \phi_{i,n-2}^+, \hat{\phi}_{i,n-1}^+ = \phi_{i,n-1}^+, \hat{\phi}_{i,1}^- = \phi_{i,1}^-, \hat{\phi}_{i,n-1}^- = \phi_{i,n-1}^-$ and counters $c_{i,0}^+ = 1, c_{i,n-2}^+ = n-1, c_{i,n-1}^+ = 1, c_{i,1}^- = n-1, c_{i,n-1}^- = 1$.

Algorithm 6 WarmUp$^+(\mathcal{N})$

1: **for** $s = 2, \ldots, n-2$ **do**
2: Draw a permutation π of $\mathcal{N}$ u.a.r.
3: **for** $k = 0, \ldots, \lfloor \frac{n}{s} \rfloor - 1$ **do**
4: $A \leftarrow \{\pi(1+ks), \ldots, \pi(s+ks)\}$
5: $v \leftarrow \nu(A)$
6: **for** $i \in A$ **do**
7: $\hat{\phi}^+_{i,s-1} \leftarrow v$
8: $c^+_{i,s-1} \leftarrow 1$
9: **end for**
10: **end for**
11: **if** $n \mod s \neq 0$ **then**
12: $A \leftarrow \{\pi(n-(n \mod s)+1), \ldots, \pi(n)\}$
13: Draw $B \in \{S \subseteq \mathcal{N} \setminus A \mid |S| = s - (n \mod s)\}$
14: $v \leftarrow \nu(A \cup B)$
15: **for** $i \in A$ **do**
16: $\hat{\phi}^+_{i,s-1} \leftarrow v$
17: $c^+_{i,s-1} \leftarrow 1$
18: **end for**
19: **end if**
20: **end for**

The warm-up for the positive Shapley subvalues iterates over all coalition sizes from 2 to $n-2$ (line 1) and draws for each size s a permutation π of $\mathcal{N}$ uniformly at random (line 2). The ordering π is sliced into coalitions of size s and each of them is used to update only the players contained in that particular coalition (lines 6–9). In particular, since each coalition A is the first to be observed for the corresponding players' stratum, the estimate $\hat{\phi}^+_{i,s-1}$ is set to the worth of A and its counter $c^+_{i,s-1}$ is set to 1. Note that for each coalition A of size s only one access to ν is made to update all affected s many players. In case that n is not a multiple of s, some players less than s are left over at the end of π (line 11). We group those with other random players to form a coalition of size s, but only update with the worth of that coalition the left out players (lines 15–18). Note that the warm-up comes without any bias, since for each player i and stratum estimate $\hat{\phi}_{i,s}$ each coalition $A \subseteq \mathcal{N}$ with $i \in A$ and $|A| = s$ has the same probability of being chosen. Finally, $\lceil \frac{n}{s} \rceil$ many coalitions are evaluated for each size s, resulting in $\sum_{s=2}^{n-2} \lceil \frac{n}{s} \rceil \in \mathcal{O}(n \log n)$ total evaluations.

Algorithm 7 WarmUp$^-(\mathcal{N})$

1: **for** $s = 2, \ldots, n-2$ **do**
2: Draw a permutation π of $\mathcal{N}$ u.a.r.
3: **for** $k = 0, \ldots, \lfloor \frac{n}{s} \rfloor - 1$ **do**
4: $A \leftarrow \{\pi(1+ks), \ldots, \pi(s+ks)\}$
5: $v \leftarrow \nu(\mathcal{N} \setminus A)$
6: **for** $i \in A$ **do**
7: $\hat{\phi}^-_{i,n-s} \leftarrow v$
8: $c^-_{i,n-s} \leftarrow 1$
9: **end for**
10: **end for**
11: **if** $n \mod s \neq 0$ **then**
12: $A \leftarrow \{\pi(n-(n \mod s)+1), \ldots, \pi(n)\}$
13: Draw $B \in \{S \subseteq \mathcal{N} \setminus A \mid |S| = s - (n \mod s)\}$ u.a.r.
14: $v \leftarrow \nu(\mathcal{N} \setminus (A \cup B))$
15: **for** $i \in A$ **do**
16: $\hat{\phi}^-_{i,n-s} \leftarrow v$
17: $c^-_{i,n-s} \leftarrow 1$
18: **end for**
19: **end if**
20: **end for**

The warm-up for the negative subvalues proceeds analogously to the previously presented positive warm-up. Instead of $\hat{\phi}^+_{i,s-1}$ and $c^+_{i,s-1}$, $\hat{\phi}^-_{i,n-s}$ and $c^-_{i,n-s}$ are updated with the wort of $\mathcal{N} \setminus A$ for all players contained in the coalition A.

B.3 Stratified SVARM$^+$

Algorithm 8 Stratified SVARM$^+$

Input: $\mathcal{N}, T \in \mathbb{N}$

1: $\hat{\phi}^+_{i,\ell}, \hat{\phi}^-_{i,\ell} \leftarrow 0$ for all $i \in \mathcal{N}$ and $\ell \in \mathcal{L}$
2: $c^+_{i,\ell}, c^-_{i,\ell} \leftarrow 0$ for all $i \in \mathcal{N}$ and $\ell \in \mathcal{L}$
3: ExactCalculation($\mathcal{N}$)
4: $t \leftarrow 2n + 1$ {consumed budget}
5: $w_s \leftarrow \tilde{P}(s)$ for all $s \in \{2, \ldots, n-2\}$ {sampling weight of size s}
6: $L_s \leftarrow \emptyset$ for all $s \in \{2, \ldots, n-2\}$ {sampled coalitions of size s}
7: $m_s \leftarrow \binom{n}{s}$ for all $s \in \{2, \ldots, n-2\}$ {reamining sets to sample of size s}
8: **while** $t < T$ and $m_s > 0$ for some s **do**
9: Draw $s_t \in \{2, \ldots, n-2\}$ with probability $\frac{w_s m_s}{\sum_{s'=2}^{n-2} w_{s'} m_{s'}}$
10: Draw A_t from $\{S \subseteq \mathcal{N} \mid |S| = s_t, S \notin L_{s_t}\}$ u.a.r.
11: Update(A_t)
12: $m_{s_t} \leftarrow m_{s_t} - 1$
13: $t \leftarrow t + 1$
14: **end while**
15: $\hat{\phi}_i \leftarrow \frac{1}{|\{c^+_{i,\ell} | \ell \in \{0,\ldots,n-1\}\}|} \sum_{\ell=0}^{n-1} \hat{\phi}^+_{i,\ell} - \frac{1}{|\{c^-_{i,\ell} | \ell \in \{0,\ldots,n-1\}\}|} \sum_{\ell=0}^{n-1} \hat{\phi}^-_{i,\ell}$ for all $i \in \mathcal{N}$

Output: $\hat{\phi}_1, \ldots, \hat{\phi}_n$

Stratified SVARM$^+$ is a modification of Stratified SVARM to deliver better empirical performance with only two slight changes that do not alter the method on a conceptual level. First, we remove the warm-up since it is less efficient in the sense that not all players estimates are updated with each sampled coalition. Hence, we only consume a budget of $2n + 1$ due to the exact calculation of the border strata before entering the main loop. Although it is extremely unlikely for a sufficiently large chosen budget T and an appropriate distribution $\tilde{P}$ over the coalition sizes, it can happen that some $c^+_{i,\ell}$ or $c^-_{i,\ell}$ are zero. In this case dividing by n the total number of strata per sign per player in line 15 would cause an unnecessary bias. Instead, we average only over all strata for which at least one sample has been observed, i.e. $c^+_{i,\ell} > 0$ respectively $c^-_{i,\ell} > 0$. Second and most important, we sample coalitions without replacement. We are aware that different ways of implementing this exist (saving substantial amounts of runtime), but we choose to demonstrate it as simply as possible. Effectively each coalition of size $s \in \{2, \ldots, n-2\}$ is assigned the weight $\frac{\tilde{P}(s)}{\binom{n}{s}}$, such that coalitions of the same size have the same weight and their weight sums up to $\tilde{P}(s)$. In each time step t a remaining coalition A_t is drawn with probability proportional to its weight (its own weight divided by the sum of all remaining coalitions' weights). We realize this by a two-step procedure: first the size s_t is drawn in line 9, then a remaining coalition of size s_t is drawn uniformly at random in line 10. For this purpose we keep track of all so far sampled coalitions of a given size s in L_s (line 6) and the number of coalitions m_s of size s left to sample (line 7). Finally, we added the condition that at least one coalition must be left to sample to the loop in line 8, in case that T is chosen larger than $2^n - 1$.

C SVARM Analysis

Notation:

- Let $\bar{T} = T - 2n$.
- Let A_i be a random set with $\mathbb{P}(A_i = S) = w_S$ for all $S \subseteq \mathcal{N}_i$.
- Let ${\sigma_i^+}^2 = \mathbb{V}[\nu(A_i \cup \{i\})]$.
- Let ${\sigma_i^-}^2 = \mathbb{V}[\nu(A_i)]$.
- Let $\bar{m}_i^+$ be number of sampled coalitions A^+ after the warm-up phase that contain i.
- Let $\bar{m}_i^-$ be number of sampled coalitions A^- after the warm-up phase that do not contain i.
- Let $m_i^+ = \bar{m}_i^+ + 1$ be total number of samples for $\hat{\phi}_i^+$.
- Let $m_i^- = \bar{m}_i^+ + 1$ be total number of samples for $\hat{\phi}_i^-$.
- Let $r_i^+ = \max_{S \subseteq \mathcal{N}_i} \nu(S \cup \{i\}) - \min_{S \subseteq \mathcal{N}_i} \nu(S \cup \{i\})$ be the range of $\nu(A_i \cup \{i\})$.
- Let $r_i^- = \max_{S \subseteq \mathcal{N}_i} \nu(S) - \min_{S \subseteq \mathcal{N}_i} \nu(S)$ be the range of $\nu(A_i)$.

Assumptions:

- $\bar{T}$ is even
- $\bar{T} > 0$

C.1 Unbiasedness of Shapley Value Estimates

To start with, we prove that the distributions P^+ and P^- are well-defined.

Lemma 1. *The distributions P^+ and P^- over $\mathcal{P}(\mathcal{N})$ are well-defined, i.e.,*

$$\sum_{S \subseteq \mathcal{N}} P^+(S) = \sum_{S \subseteq \mathcal{N}} P^-(S) = 1.$$

Proof. The statement is easily shown for P^+ by grouping the coalitions by size. We derive:

$$\begin{aligned}
&\sum_{S \subseteq \mathcal{N}} P^+(S) \\
&= \sum_{\ell=1}^{n} \sum_{\substack{S \subseteq \mathcal{N} \\ |S|=\ell}} P^+(S) \\
&= \sum_{\ell=1}^{n} \sum_{\substack{S \subseteq \mathcal{N} \\ |S|=\ell}} \frac{1}{\ell \binom{n}{\ell} H_n} \\
&= \frac{1}{H_n} \sum_{\ell=1}^{n} \frac{1}{\ell} \\
&= 1.
\end{aligned}$$

One can prove the desired property analogously for P^-. □

For the remainder of this section we assume that $T \geq 2n + 1$ such that the warm-up phase can be completed by SVARM.

Lemma 2. *For each player $i \in \mathcal{N}$ the positive and negative Shapley Value estimates $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ are unbiased, i.e.,*

$$\mathbb{E}\left[\hat{\phi}_i^+\right] = \phi_i^+ \quad \textit{and} \quad \mathbb{E}\left[\hat{\phi}_i^-\right] = \phi_i^- .$$

Proof. Let $\bar{m}_i^+$ be the number of coalitions sampled after the warm-up phase that contain i and m_i^+ be the total number of samples used to update $\hat{\phi}_i^+$, thus $m_i^+ = \bar{m}_i^+ + 1$. Further, let A_m^+ for $m \in \{1, 3, 5, \ldots, T-1\}$ be the sampled coalitions for updating the positive Shapley values $(\hat{\phi}_i^+)_{i \in \mathcal{N}}$, then we can write the positive Shapley value of player $i \in \mathcal{N}$ as

$$\begin{aligned} \hat{\phi}_i^+ &= \frac{1}{m_i^+} \sum_{\tilde{m}=1}^{T/2} \nu(A_{2\tilde{m}-1}^+) \mathbb{I}_{\{i \in A_{2\tilde{m}-1}^+\}} \\ &= \frac{1}{m_i^+} \left(\nu(A_{2i-1}^+) + \sum_{\tilde{m}=n}^{T/2} \nu(A_{2\tilde{m}-1}^+) \mathbb{I}_{\{i \in A_{2\tilde{m}-1}^+\}} \right), \end{aligned} \tag{19}$$

where $\mathbb{I}_{\cdot}$ denotes the indicator function, and we used that during the warm-up phase ($m \leq 2n$) there is for each player i only one A_m^+ to update the corresponding positive Shapley value, namely at time step $2i-1$. First, we show for each odd $m \geq 2n$ and $S \subseteq \mathcal{N}_i$ that $\mathbb{P}(A_m^+ = S \cup \{i\} \mid i \in A_m^+) = w_S$. Note that since $A_m^+ \sim P^+$ (see (6)) it holds that

$$\begin{aligned} \mathbb{P}(i \in A_m^+) &= \sum_{\ell=1}^{n} \mathbb{P}(i \in A_m^+, |A_m^+| = \ell) \\ &= \sum_{\ell=1}^{n} \mathbb{P}(i \in A_m^+ \mid |A_m^+| = \ell) \cdot \mathbb{P}(|A_m^+| = \ell) \\ &= \sum_{\ell=1}^{n} \frac{\ell}{n} \cdot \frac{1}{\ell \cdot H_{n-1}} \\ &= \frac{1}{H_n}. \end{aligned} \tag{20}$$

With this, we derive

$$\begin{aligned} \mathbb{P}(A_m^+ = S \cup \{i\} \mid i \in A_m^+) &= \frac{\mathbb{P}(A_m^+ = S \cup \{i\})}{\mathbb{P}(i \in A_i^+)} \\ &= H_n \cdot P^+(S \cup \{i\}) \\ &= \frac{1}{(|S+1|) \cdot \binom{n}{|S+1|}} \\ &= w_S. \end{aligned}$$

Since $A_{2i-1}^+ \setminus \{i\} \sim P^w$ it holds that $\mathbb{P}(A_{2i-1}^+ \setminus \{i\} = S) = w_S$ for any $i \in \mathcal{N}$ and $S \subset \mathcal{N}_i$.
Taking all of this into account, we derive for $\hat{\phi}_i^+$ using (19):

$$\begin{aligned} \mathbb{E}\left[\hat{\phi}_i^+ \mid m_i^+\right] &= \frac{1}{m_i^+} \left(\mathbb{E}\left[\nu(A_{2i-1}^+) \mid m_i^+\right] + \mathbb{E}\left[\sum_{\tilde{m}=n}^{T/2} \nu(A_{2\tilde{m}-1}^+) \mathbb{I}_{\{i \in A_{2\tilde{m}-1}^+\}} \mid m_i^+ \right] \right) \\ &= \frac{1}{m_i^+} \Big(\sum_{S \subseteq \mathcal{N}_i} \mathbb{P}(A_{2i-1}^+ \setminus \{i\} = S) \cdot \nu(S \cup \{i\}) \\ &\qquad + \mathbb{E}\Big[\sum_{\tilde{m}=n}^{T/2} \mathbb{I}_{\{i \in A_{2\tilde{m}-1}^+\}} \sum_{S \subset \mathcal{N}_i} \mathbb{P}(A_{2\tilde{m}-1}^+ = S \cup \{i\} \mid i \in A_{2\tilde{m}-1}^+) \cdot \nu(S \cup \{i\}) \mid m_i^+ \Big] \Big) \\ &= \frac{1}{m_i^+} \left(\sum_{S \subseteq \mathcal{N}_i} w_S \cdot \nu(S \cup \{i\}) + \bar{m}_i^+ \sum_{S \subseteq \mathcal{N}_i} w_S \cdot \nu(S \cup \{i\}) \right) \\ &= \phi_i^+. \end{aligned}$$

Finally, we conclude:

$$\begin{aligned}\mathbb{E}\left[\hat{\phi}_i^+\right] &= \sum_{m=1}^{\frac{\bar{T}}{2}} \mathbb{E}\left[\hat{\phi}_i^+ \mid m_i^+ = m\right] \cdot \mathbb{P}(m_i^+ = m) \\ &= \sum_{m=1}^{\frac{\bar{T}}{2}} \phi_i^+ \cdot \mathbb{P}(m_i^+ = m) \\ &= \phi_i^+.\end{aligned}$$

Analogously we derive $\mathbb{E}\left[\hat{\phi}_i^-\right] = \phi_i^-$ by defining $\bar{m}_i^-$, m_i^-, and A_m^- for $m \in \{2, 4, 6, \ldots, T\}$ similarly as for their positive counterparts. □

Theorem 1 *For each player $i \in \mathcal{N}$ the estimate $\hat{\phi}_i$ obtained by SVARM is unbiased, i.e.,*

$$\mathbb{E}\left[\hat{\phi}_i\right] = \phi_i.$$

Proof. We apply Lemma 2 and obtain in combination with Equation (4):

$$\begin{aligned}\mathbb{E}\left[\hat{\phi}_i\right] &= \mathbb{E}\left[\hat{\phi}_i^+\right] - \mathbb{E}\left[\hat{\phi}_i^-\right] \\ &= \phi_i^+ - \phi_i^- \\ &= \phi_i.\end{aligned}$$

□

C.2 Sample Numbers

Lemma 3. *For any $i \in \mathcal{N}$ the expected number of updates of $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ after the warm-up phase is*

$$\mathbb{E}\left[\bar{m}_i^+\right] = \mathbb{E}\left[\bar{m}_i^-\right] = \frac{\bar{T}}{2H_n}.$$

Proof. First, we observe that $\bar{m}_i^+$ is binomially distributed with $\bar{m}_i^+ \sim Bin\left(\frac{\bar{T}}{2}, \frac{1}{H_n}\right)$ because $\frac{\bar{T}}{2}$ many pairs are sampled and each independently sampled coalition A^+ contains the player i with probability H_n^{-1}, see (20).
Consequently, we obtain

$$\mathbb{E}\left[\bar{m}_i^+\right] = \frac{\bar{T}}{2H_n}.$$

Similarly, we observe that m_i^- is also binomially distributed with $m_i^- \sim Bin\left(\frac{\bar{T}}{2}, \frac{1}{H_n}\right)$, leading to the same expected number of updates. □

C.3 Variance and Squared Error

Lemma 4. *The variance of any player's Shapley value estimate $\hat{\phi}_i$ given the number of samples m_i^+ and m_i^- is exactly*

$$\mathbb{V}\left[\hat{\phi}_i \mid m_i^+, m_i^-\right] = \frac{{\sigma_i^+}^2}{m_i^+} + \frac{{\sigma_i^-}^2}{m_i^-}.$$

Proof. We first decompose the variance of $\hat{\phi}_i$ into the variances of $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ and their covariance:

$$\mathbb{V}\left[\hat{\phi}_i \mid m_i^+, m_i^-\right] = \left(\mathbb{V}\left[\hat{\phi}_i^+ \mid m_i^+\right] + \mathbb{V}\left[\hat{\phi}_i^- \mid m_i^-\right] - 2\mathrm{Cov}\left(\hat{\phi}_i^+, \hat{\phi}_i^- \mid m_i^+, m_i^-\right)\right)$$

We derive for the variance of $\hat{\phi}_i^+$:

$$\begin{aligned}\mathbb{V}\left[\hat{\phi}_i^+ \mid m_i^+\right] &= \frac{1}{{m_i^+}^2}\sum_{m=0}^{\bar{m}_i^+}\mathbb{V}\left[\nu(A_{i,m}^+)\right] \\ &= \frac{{\sigma_i^+}^2}{m_i^+}.\end{aligned}$$

Similarly we obtain for $\hat{\phi}_i^-$:

$$\mathbb{V}\left[\hat{\phi}_i^- \mid m_i^-\right] = \frac{{\sigma_i^-}^2}{m_i^-}.$$

The covariance of $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ is zero because both are updated with sampled coalitions drawn independently of each other. Thus, we conclude:

$$\mathbb{V}\left[\hat{\phi}_i \mid m_i^+, m_i^-\right] = \frac{{\sigma_i^+}^2}{m_i^+} + \frac{{\sigma_i^-}^2}{m_i^-}.$$

□

Lemma 5. *For the sample numbers of any player $i \in \mathcal{N}$ holds*

$$\mathbb{E}\left[\frac{1}{m_i^+}\right] = \mathbb{E}\left[\frac{1}{m_i^-}\right] \leq \frac{2H_n}{\bar{T}}.$$

Proof. By combining Equation (3.4) in Chao and Strawderman [1972]:

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1-(1-p)^{m+1}}{(m+1)p} \leq \frac{1}{mp} = \frac{1}{\mathbb{E}[X]},$$

for any binomially distributed random variable $X \sim Bin(m,p)$ with Lemma 3, we obtain:

$$\mathbb{E}\left[\frac{1}{m_i^+}\right] = \mathbb{E}\left[\frac{1}{1+\bar{m}_i^+}\right] \leq \frac{1}{\mathbb{E}\left[\bar{m}_i^+\right]} = \frac{2H_n}{\bar{T}}.$$

Notice that m_i^+ and m_i^- are identically distributed. □

Theorem 2 *The variance of any player's Shapley value estimate $\hat{\phi}_i$ is bounded by*

$$\mathbb{V}\left[\hat{\phi}_i\right] \leq \frac{2H_n}{\bar{T}}\left({\sigma_i^+}^2 + {\sigma_i^-}^2\right).$$

Proof. The combination of Lemma 4 and Lemma 5 yields:

$$\begin{aligned}\mathbb{V}\left[\hat{\phi}_i\right] &= \mathbb{E}\left[\mathbb{V}\left[\hat{\phi}_i \mid m_i^+, m_i^-\right]\right] \\ &= \mathbb{E}\left[\frac{{\sigma_i^+}^2}{m_i^+} + \frac{{\sigma_i^-}^2}{m_i^-}\right] \\ &= {\sigma_i^+}^2\mathbb{E}\left[\frac{1}{m_i^+}\right] + {\sigma_i^-}^2\left[\frac{1}{m_i^-}\right] \\ &\leq \frac{2H_n}{\bar{T}}\left({\sigma_i^+}^2 + {\sigma_i^-}^2\right).\end{aligned}$$

□

Corollary 1 *The expected squared error of any player's Shapley value estimate is bounded by*

$$\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \leq \frac{2H_n}{\bar{T}}\left({\sigma_i^+}^2 + {\sigma_i^-}^2\right).$$

Proof. The bias-variance decomposotion allows us to plug in the unbiasedness of $\hat{\phi}_i$ shown in Theorem 1 and the bound on the variance from

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] &= \left(\mathbb{E}\left[\hat{\phi}_i\right] - \phi_i\right)^2 + \mathbb{V}\left[\hat{\phi}_i\right] \\ &\leq \frac{2H_n}{\bar{T}}\left({\sigma_i^+}^2 + {\sigma_i^-}^2\right).\end{aligned}$$

□

C.4 Probabilistic Bounds

Theorem 3 *Fix any player $i \in \mathcal{N}$ and $\varepsilon > 0$. The probability that the Shapley value estimate $\hat{\phi}_i$ deviates from ϕ_i by a margin of ε or greater is bounded by*

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq \frac{2H_n}{\varepsilon^2 \bar{T}}\left({\sigma_i^-}^2 + {\sigma_i^+}^2\right).$$

Proof. The bound on the variance of $\hat{\phi}_i$ in Theorem 2 allows us to apply Chebyshev's inequality:

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq \frac{\mathbb{V}\left[\hat{\phi}_i\right]}{\varepsilon^2} \leq \frac{2H_n}{\varepsilon^2 \bar{T}}\left({\sigma_i^-}^2 + {\sigma_i^+}^2\right).$$

□

Corollary 3. *Fix any player $i \in \mathcal{N}$ and $\delta \in (0, 1]$. The Shapley value estimate $\hat{\phi}_i$ deviates from ϕ_i by a margin of ε or greater with probability not greater than δ, i.e.,*

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq \delta \quad \textit{for} \quad \varepsilon = \sqrt{\frac{2H_n}{\delta \bar{T}}\left({\sigma_i^+}^2 + {\sigma_i^-}^2\right)}.$$

Lemma 6. *For any fixed player $i \in \mathcal{N}$ and $\varepsilon > 0$ holds*

$$\mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+\right) \leq 2\exp\left(-\frac{2m_i^+\varepsilon^2}{{r_i^+}^2}\right) \quad \textit{and} \quad \mathbb{P}\left(|\hat{\phi}_i^- - \phi_i^-| \geq \varepsilon \mid m_i^-\right) \leq 2\exp\left(-\frac{2m_i^+\varepsilon^2}{{r_i^+}^2}\right).$$

Proof. We prove the statement for $\hat{\phi}_i$ by making use of Hoeffding's inequality in combination with the unbiasedness of the positive and negative Shapley value estimates shown in Lemma 2. The proof for $\hat{\phi}_i^-$ is analogous.

$$\begin{aligned}\mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+\right) &= \mathbb{P}\left(|\hat{\phi}_i^+ - \mathbb{E}\left[\hat{\phi}_i+\right]| \mid m_i^+\right) \\ &= \mathbb{P}\left(\left|\sum_{m=0}^{\bar{m}_i^+} \nu(A_{i,m}^+) - \mathbb{E}\left[\sum_{m=0}^{\bar{m}_i^+} \nu(A_{i,m}^+)\right]\right| \geq m_i^+\varepsilon \mid m_i^+\right) \\ &\leq 2\exp\left(-\frac{2m_i^+\varepsilon^2}{{r_i^+}^2}\right).\end{aligned}$$

□

Lemma 7. *For any fixed player $i \in \mathcal{N}$ and $\varepsilon > 0$ holds:*

- $\mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon\right) \leq \exp\left(-\frac{\bar{T}}{4{H_n}^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_i^+}^2}\right)^{\left\lfloor\frac{\bar{T}}{4H_n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_i^+}^2}\right)-1}$,
- $\mathbb{P}\left(|\hat{\phi}_i^- - \phi_i^-| \geq \varepsilon\right) \leq \exp\left(-\frac{\bar{T}}{4{H_n}^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_i^-}^2}\right)^{\left\lfloor\frac{\bar{T}}{4H_n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_i^-}^2}\right)-1}$.

Proof. We prove the statement for $\hat{\phi}_i^+$. The proof for $\hat{\phi}_i^-$ is analogous. To begin with, we derive with the help of Hoeffding's inequality for binomial distributions a bound for the probability of $\bar{m}_i^+$ not exceeding $\frac{\bar{T}}{4H_n}$:

$$\begin{aligned}
\mathbb{P}\left(\bar{m}_i^+ \leq \frac{\bar{T}}{4H_n}\right) &\leq \mathbb{P}\left(\mathbb{E}\left[\bar{m}_i^+\right] - \bar{m}_i^+ \geq \mathbb{E}\left[\bar{m}_i^+\right] - \frac{\bar{T}}{4H_n}\right) \\
&\leq \exp\left(-\frac{4\left(\mathbb{E}\left[\bar{m}_i^+\right] - \frac{\bar{T}}{4H_n}\right)^2}{\bar{T}}\right) \\
&\leq \exp\left(-\frac{\bar{T}}{4{H_n}^2}\right),
\end{aligned}$$

where we used the lower bound on $\mathbb{E}\left[\bar{m}_i^+\right]$ shown in Lemma 3. Next, we derive with the help of Lemma 6 a statement of technical nature to be used later:

$$\begin{aligned}
&\sum_{m=\lfloor\frac{\bar{T}}{4H_n}\rfloor+1}^{\frac{\bar{T}}{2}} \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+ = m\right) \\
&\leq 2 \sum_{m=\lfloor\frac{\bar{T}}{4H_n}\rfloor+1}^{\frac{\bar{T}}{2}} \exp\left(-\frac{2m\varepsilon^2}{{r_i^+}^2}\right) \\
&= 2\sum_{m=0}^{\frac{\bar{T}}{2}} \exp\left(-\frac{2\varepsilon^2}{{r_i^+}^2}\right)^m - 2\sum_{m=0}^{\lfloor\frac{\bar{T}}{4H_n}\rfloor} \exp\left(-\frac{2\varepsilon^2}{{r_i^+}^2}\right)^m \\
&= 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_i^+}^2}\right)^{\lfloor\frac{\bar{T}}{4H_n}\rfloor} - \exp\left(-\frac{\varepsilon^2}{{r_i^+}^2}\right)^{\bar{T}}}{\exp\left(\frac{2\varepsilon^2}{{r_i^+}^2}\right) - 1} \\
&\leq 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_i^+}^2}\right)^{\lfloor\frac{\bar{T}}{4H_n}\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_i^+}^2}\right) - 1}.
\end{aligned}$$

At last, putting both findings together, we derive our claim:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon\right) \\
&= \sum_{m=1}^{\frac{\bar{T}}{2}} \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+ = m\right) \cdot \mathbb{P}\left(m_i^+ = m\right) \\
&= \sum_{m=1}^{\lfloor\frac{\bar{T}}{4H_n}\rfloor} \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+ = m\right) \cdot \mathbb{P}\left(m_i^+ = m\right) + \sum_{m=\lfloor\frac{\bar{T}}{4H_n}\rfloor+1}^{\frac{\bar{T}}{2}} \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+ = m\right) \cdot \mathbb{P}\left(m_i^+ = m\right) \\
&\leq \mathbb{P}\left(\bar{m}_i^+ \leq \left\lfloor\frac{\bar{T}}{4H_n}\right\rfloor\right) + \sum_{m=\lfloor\frac{\bar{T}}{4H_n}\rfloor+1}^{\frac{\bar{T}}{2}} \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon \mid m_i^+ = m\right) \\
&\leq \exp\left(-\frac{\bar{T}}{4{H_n}^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_i^+}^2}\right)^{\lfloor\frac{\bar{T}}{4H_n}\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_i^+}^2}\right) - 1}.
\end{aligned}$$

□

Theorem 4 *For any fixed player $i \in \mathcal{N}$ and $\varepsilon > 0$ the probability that the Shapley value estimate $\hat{\phi}_i$ deviates from ϕ_i by a margin of ε or greater is bounded by*

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\bar{T}}{4{H_n}^2}\right) + 4\frac{\exp\left(-\frac{2\varepsilon^2}{(r_i^+ + r_i^-)^2}\right)^{\left\lfloor \frac{\bar{T}}{4H_n} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{(r_i^+ + r_i^-)^2}\right) - 1}.$$

Proof.

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \\
&= \mathbb{P}\left(|(\hat{\phi}_i^+ - \phi_i^+) + (\phi_i^- - \hat{\phi}_i^-)| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| + |\hat{\phi}_i^- - \phi_i^-| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \frac{\varepsilon r_i^+}{r_i^+ + r_i^-}\right) + \mathbb{P}\left(|\hat{\phi}_i^- - \phi_i^-| \geq \frac{\varepsilon r_i^-}{r_i^+ + r_i^-}\right) \\
&\leq 2\exp\left(-\frac{\bar{T}}{4{H_n}^2}\right) + 4\frac{\exp\left(-\frac{2\varepsilon^2}{(r_i^+ + r_i^-)^2}\right)^{\left\lfloor \frac{\bar{T}}{4H_n} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{(r_i^+ + r_i^-)^2}\right) - 1}.
\end{aligned}$$

□

D Stratified SVARM Analysis

Notation:

- Let $\mathcal{L} = \{0, \dots, n-1\}$, $\mathcal{L}^+ = \{1, \dots, n-3\}$, and $\mathcal{L}^- = \{2, \dots, n-2\}$.
- Let $W = 2n + 1 + 2\sum_{s=2}^{n-2} \lceil \frac{n}{s} \rceil$ denote the length of the warm-up phase.
- Let $\bar{T} = T - W$ be the available steps after the warm-up phase.
- Let $m^+_{i,\ell} = \#\{t \mid i \in A_t, |A_t| = \ell + 1\}$ be the total number of samples used to update $\hat{\phi}^+_{i,\ell}$.
- Let $m^-_{i,\ell} = \#\{t \mid i \notin A_t, |A_t| = \ell\}$ be the total number of samples used to update $\hat{\phi}^-_{i,\ell}$.
- Let $\bar{m}^+_{i,\ell} = \#\{t \mid i \in A_t, |A_t| = \ell + 1, t > W\}$ be the number of samples used to update $\hat{\phi}^+_{i,\ell}$ after the warm-up phase.
- Let $\bar{m}^-_{i,\ell} = \#\{t \mid i \notin A_t, |A_t| = \ell, t > W\}$ be the number of samples used to update $\hat{\phi}^-_{i,\ell}$ after the warm-up phase.
- Let $A^+_{i,\ell,k}$ be the k-th set used to update $\phi^+_{i,\ell}$ and $A^-_{i,\ell,k}$ the k-th set used to update $\phi^-_{i,\ell}$.
- Let $A_{i,\ell}$ be a random set with $\mathbb{P}(A_{i,\ell} = S) = \frac{1}{\binom{n-1}{\ell}}$ for all $S \subseteq \mathcal{N} \setminus \{i\}$ with $|S| = \ell$.
- Let $\hat{\phi}^+_{i,\ell} = \frac{1}{m^+_{i,\ell}} \sum_{k=1}^{m^+_{i,\ell}} \nu(A^+_{i,\ell,k})$ and $\hat{\phi}^-_{i,\ell} = \frac{1}{m^-_{i,\ell}} \sum_{k=1}^{m^-_{i,\ell}} \nu(A^-_{i,\ell,k})$.
- Let $\hat{\phi}_i = \frac{1}{n} \sum_{\ell=0}^{n-1} \hat{\phi}^+_{i,\ell} - \hat{\phi}^-_{i,\ell}$.
- Let ${\sigma^+_{i,\ell}}^2 = \mathbb{V}\left[\nu(A_{i,\ell} \cup \{i\})\right]$ and ${\sigma^-_{i,\ell}}^2 = \mathbb{V}\left[\nu(A_{i,\ell})\right]$.
- Let $r^+_{i,\ell} = \max_{S \subseteq \mathcal{N} : i \notin S, |S| = \ell} \nu(S \cup \{i\}) - \min_{S \subseteq \mathcal{N} : i \notin S, |S| = \ell} \nu(S \cup \{i\})$ be the range of $\nu(A^+_{i,\ell,k})$.
- Let $r^-_{i,\ell} = \max_{S \subseteq \mathcal{N} : i \notin S, |S| = \ell} \nu(S) - \min_{S \subseteq \mathcal{N} : i \notin S, |S| = \ell} \nu(S)$ be the range of $\nu(A^-_{i,\ell,k})$.
- Let $R^+_i = \sum_{\ell=1}^{n-3} r^+_{i,\ell}$ and $R^-_i = \sum_{\ell=2}^{n-2} r^-_{i,\ell}$.

Assumptions:

- $n \geq 4$, for $n \leq 3$ the algorithm computes all Shapley values exactly.

D.1 Unbiasedness of Shapley Value Estimates

Lemma 8. *Due to the exact calculation, the following estimates are exact for all $i \in \mathcal{N}$:*

- $\hat{\phi}^+_{i,0} = \phi^+_{i,0} = \nu(\{i\})$
- $\hat{\phi}^+_{i,n-2} = \phi^+_{i,n-2} = \frac{1}{n-1} \sum_{j \in \mathcal{N} : j \neq i} \nu(\mathcal{N} \setminus \{j\})$
- $\hat{\phi}^+_{i,n-1} = \phi^+_{i,n-1} = \nu(\mathcal{N})$
- $\hat{\phi}^-_{i,0} = \phi^-_{i,0} = \nu(\emptyset) = 0$
- $\hat{\phi}^-_{i,1} = \phi^-_{i,1} = \frac{1}{n-1} \sum_{j \in \mathcal{N} : j \neq i} \nu(\{j\})$
- $\hat{\phi}^-_{i,n-1} = \phi^-_{i,n-1} = \nu(\mathcal{N} \setminus \{i\})$

Lemma 9. *All remaining estimates that are not calculated exactly are unbiased, i.e., for all $i \in \mathcal{N}$:*

- $\mathbb{E}\left[\hat{\phi}_{i,\ell}^{+}\right] = \phi_{i,\ell}^{+}$ *for all* $\ell \in \mathcal{L}^{+}$
- $\mathbb{E}\left[\hat{\phi}_{i,\ell}^{-}\right] = \phi_{i,\ell}^{-}$ *for all* $\ell \in \mathcal{L}^{-}$

Proof. We prove the statement only for $\hat{\phi}_{i,\ell}^{+}$ as the proof for $\hat{\phi}_{i,\ell}^{-}$ is analogous. Fix any $i \in \mathcal{N}$ and $\ell \in \mathcal{L}^{+}$. As soon as the size s_t of the to be sampled coalition $|A_t|$ is fixed, A_t is sampled uniformly from $\{S \subseteq \mathcal{N} \mid |S| = s_t\}$. This allows us to state for every A_t and any $S \subseteq \mathcal{N}$ with $|S| = \ell + 1$ and $i \notin S$:

$$\mathbb{P}\left(A_t = S \cup \{i\} \mid i \in A_t, |A_t| = \ell + 1\right) = \frac{1}{\binom{n-1}{\ell}}.$$

Continuing, we derive for the expectation of $\hat{\phi}_{i,\ell}^{+}$ given the number of samples $m_{i,\ell}^{+}$:

$$\begin{aligned}
&\mathbb{E}\left[\hat{\phi}_{i,\ell}^{+} \mid m_{i,\ell}^{+}\right] \\
&= \mathbb{E}\left[\frac{1}{m_{i,\ell}^{+}} \sum_{k=1}^{m_{i,\ell}^{+}} \nu(A_{i,\ell,k}^{+}) \middle| m_{i,\ell}^{+}\right] \\
&= \frac{1}{m_{i,\ell}^{+}} \sum_{k=1}^{m_{i,\ell}^{+}} \mathbb{E}\left[\nu(A_{i,\ell,k}^{+}) \mid m_{i,\ell}^{+}\right] \\
&= \frac{1}{m_{i,\ell}^{+}} \sum_{k=1}^{m_{i,\ell}^{+}} \sum_{S \subseteq \mathcal{N} \setminus \{i\} : |S| = \ell} \mathbb{P}\left(A_{i,\ell,k}^{+} = S \cup \{i\} \mid i \in A_{i,\ell,k}^{+}, |A_{i,\ell,k}^{+}| = \ell + 1\right) \cdot \nu(S \cup \{i\}) \\
&= \frac{1}{m_{i,\ell}^{+}} \sum_{k=1}^{m_{i,\ell}^{+}} \sum_{S \subseteq \mathcal{N} \setminus \{i\} : |S| = \ell} \frac{1}{\binom{n-1}{\ell}} \cdot \nu(S \cup \{i\}) \\
&= \frac{1}{m_{i,\ell}^{+}} \sum_{k=1}^{m_{i,\ell}^{+}} \phi_{i,\ell}^{+} \\
&= \phi_{i,\ell}^{+}.
\end{aligned}$$

Note that the term is well defined, since $m_{i,\ell}^{+} \in \{1, \ldots, T\}$ due to the warm-up phase. We conclude:

$$\begin{aligned}
\mathbb{E}\left[\hat{\phi}_{i,\ell}^{+}\right] &= \sum_{m=1}^{T} \mathbb{E}\left[\hat{\phi}_{i,\ell}^{+} \mid m_{i,\ell}^{+} = m\right] \cdot \mathbb{P}\left(m_{i,\ell}^{+} = m\right) \\
&= \sum_{m=1}^{T} \phi_{i,\ell}^{+} \cdot \mathbb{P}\left(m_{i,\ell}^{+} = m\right) \\
&= \phi_{i,\ell}^{+}.
\end{aligned}$$

□

Theorem 5 *The Shapley value estimates for all* $i \in \mathcal{N}$ *are unbiased, i.e.,*

$$\mathbb{E}\left[\hat{\phi}_i\right] = \phi_i.$$

Proof. By applying Lemma 8 and Lemma 9 we obtain:

$$\begin{aligned}
\mathbb{E}\left[\hat{\phi}_i\right] &= \frac{1}{n} \sum_{\ell=0}^{n-1} \mathbb{E}\left[\hat{\phi}_{i,\ell}^{+}\right] - \mathbb{E}\left[\hat{\phi}_{i,\ell}^{-}\right] \\
&= \frac{1}{n} \sum_{\ell=0}^{n-1} \phi_{i,\ell}^{+} - \phi_{i,\ell}^{-} \\
&= \phi_i.
\end{aligned}$$

□

D.2 Sample numbers

Lemma 10. *For any $i \in \mathcal{N}$ the numer of updates $\bar{m}^+_{i,\ell}$ and $m^-_{i,\ell}$ are binomially distributed with*

$$\bar{m}^+_{i,\ell} \sim Bin\left(\bar{T}, \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1)\right) \text{ for all } \ell \in \mathcal{L}^+$$
$$\text{and} \quad \bar{m}^-_{i,\ell} \sim Bin\left(\bar{T}, \frac{n-\ell}{n} \cdot \tilde{P}(\ell)\right) \text{ for all } \ell \in \mathcal{L}^-.$$

Proof. We argue that there are $\bar{T}$ many independent time steps in which $\hat{\phi}^+_{i,\ell}$ can be updated. If $|A_t| = \ell + 1$ then i is included in A_t with a probability of $\frac{\ell+1}{n}$ due to the uniform sampling of A_t given that its size s_t is fixed, leading to an update. Since the choice of size and members of the set A_t are independent, the probability of $\hat{\phi}^+_{i,\ell}$ being updated in time step t is $\frac{\ell+1}{n} \cdot \tilde{P}(\ell+1)$. The same argument holds true for $\hat{\phi}^-_{i,\ell}$ with an update probability of $\frac{n-\ell}{n} \cdot \tilde{P}(\ell)$ in each time step. □

Lemma 11. *For any $i \in \mathcal{N}$ the expected number of updates of $\hat{\phi}^+_{i,\ell}$ and $\hat{\phi}^-_{i,\ell}$ after the warm-up phase is at least*

$$\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] \geq \frac{\bar{T}}{2n \log n} \text{ for all } \ell \in \mathcal{L}^+$$
$$\text{and} \quad \mathbb{E}\left[\bar{m}^-_{i,\ell}\right] \geq \frac{\bar{T}}{2n \log n} \text{ for all } \ell \in \mathcal{L}^-.$$

Proof. In the following we distinguish between different cases, depending on the parity of n and size of ℓ. We will use the bound $H_n \leq \log n + 1$ and the following inequalities multiple times which hold true for $n \geq 4$:

$$\frac{1}{H_{\frac{n-1}{2}} - 1} \geq \frac{1}{\log n} \quad \text{and} \quad \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \geq 1.$$

We begin with the case of $n \nmid 2$ and $\ell \leq \frac{n-1}{2} - 1$:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n} \cdot \frac{1}{H_{\frac{n-1}{2}} - 1} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n} \cdot \frac{n-\ell}{\ell} \cdot \frac{1}{H_{\frac{n-1}{2}} - 1} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned}$$

For $n \nmid 2$ and $\ell = \frac{n-1}{2}$ we obtain:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n} \cdot \frac{\ell+1}{n-\ell-1} \cdot \frac{1}{H_{\frac{n-1}{2}} - 1} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n} \cdot \frac{n-\ell}{\ell} \cdot \frac{1}{H_{\frac{n-1}{2}} - 1} \\ &\geq \frac{\bar{T}}{2n \log n}.\end{aligned}$$

For $n \nmid 2$ and $\ell \geq \frac{n+1}{2}$ we obtain:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n} \cdot \frac{\ell+1}{n-\ell-1} \cdot \frac{1}{H_{\frac{n-1}{2}} - 1} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n} \cdot \frac{1}{H_{\frac{n-1}{2}} - 1} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned}$$

Switching to $n \mid 2$, we start with $\ell = \frac{n}{2} - 1$:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n \log n} \cdot \frac{n-\ell}{\ell} \cdot \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned}$$

For $n \mid 2$ and $\ell = \frac{n}{2}$ we derive:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n \log n} \cdot \frac{\ell+1}{n-\ell-1} \cdot \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n \log n}\end{aligned}$$

For $n \mid 2$ and $\ell \leq \frac{n}{2} - 2$ we derive:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n \log n} \cdot \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n \log n} \cdot \frac{n-\ell}{\ell} \cdot \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned}$$

Finally, $n \mid 2$ and $\ell \geq \frac{n}{2} + 1$ yields:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^+_{i,\ell}\right] &= \bar{T} \cdot \frac{\ell+1}{n} \cdot \tilde{P}(\ell+1) \\ &= \frac{\bar{T}}{2n \log n} \cdot \frac{\ell+1}{n-\ell-1} \cdot \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned} \qquad \begin{aligned}\mathbb{E}\left[\bar{m}^-_{i,\ell}\right] &= \bar{T} \cdot \frac{n-\ell}{n} \cdot \tilde{P}(\ell) \\ &= \frac{\bar{T}}{2n \log n} \cdot \frac{n \log n - 1}{n\left(H_{\frac{n}{2}-1} - 1\right)} \\ &\geq \frac{\bar{T}}{2n \log n}\end{aligned}$$

□

D.3 Variance and Expected Squared Error

Lemma 12. *The variance of any player's Shapley value estimate $\hat{\phi}_i$ given the number of samples $m^+_{i,\ell}$ and $m^-_{i,\ell}$ for all $\ell \in \mathcal{L}$ is given by*

$$\mathbb{V}\left[\hat{\phi}_i \middle| \left(m^+_{i,\ell}\right)_{\ell \in \mathcal{L}^+}, \left(m^-_{i,\ell}\right)_{\ell \in \mathcal{L}^-}\right] = \frac{1}{n^2} \sum_{\ell=1}^{n-3} \frac{{\sigma^+_{i,\ell}}^2}{m^+_{i,\ell}} + \frac{{\sigma^-_{i,\ell+1}}^2}{m^-_{i,\ell+1}}.$$

Proof. We first decompose the variance of $\hat{\phi}_i$ into the variances of $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ and their covariance:

$$\begin{aligned}
&\mathbb{V}\left[\hat{\phi}_i \mid \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}, \left(m_{i,\ell}^-\right)_{\ell\in\mathcal{L}^-}\right] \\
&= \mathbb{V}\left[\hat{\phi}_i^+ \mid \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}\right] + \mathbb{V}\left[\hat{\phi}_i^- \mid \left(m_{i,\ell}^-\right)_{\ell\in\mathcal{L}^-}\right] - 2\text{Cov}\left(\hat{\phi}_i^+, \hat{\phi}_i^- \mid \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}, \left(m_{i,\ell}^-\right)_{\ell\in\mathcal{L}^-}\right) \\
&= \mathbb{V}\left[\hat{\phi}_i^+ \mid \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}\right] + \mathbb{V}\left[\hat{\phi}_i^- \mid \left(m_{i,\ell}^-\right)_{\ell\in\mathcal{L}^-}\right].
\end{aligned}$$

where we used the observation that $\hat{\phi}_i^+$ and $\hat{\phi}_i^-$ are independent. We derive for $\hat{\phi}_i^+$:

$$\begin{aligned}
\mathbb{V}\left[\hat{\phi}_i^+ \mid \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}\right] &= \frac{1}{n^2}\sum_{\ell=1}^{n-3}\mathbb{V}\left[\hat{\phi}_{i,\ell}^+ \mid m_{i,\ell}^+\right] + \sum_{\ell\neq\ell'}\text{Cov}\left(\hat{\phi}_{i,\ell}^+, \hat{\phi}_{i,\ell'}^+ \mid m_{i,\ell}^+, m_{i,\ell'}^+\right) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{n-3}\mathbb{V}\left[\hat{\phi}_{i,\ell}^+ \mid m_{i,\ell}^+\right] \\
&= \frac{1}{n^2}\sum_{\ell=1}^{n-3}\frac{{\sigma_{i,\ell}^+}^2}{m_{i,\ell}^+},
\end{aligned}$$

where we used the observation that $\hat{\phi}_{i,\ell}^+$ and $\hat{\phi}_{i,\ell'}^+$ are independent for $\ell \neq \ell'$. Note that $\hat{\phi}_{i,0}^+$, $\hat{\phi}_{i,n-2}^+$, $\hat{\phi}_{i,n-1}^+$, $\hat{\phi}_{i,0}^-$, $\hat{\phi}_{i,1}^-$, and $\hat{\phi}_{i,n-1}^-$ are constants without variance. A similar result can be obtained for $\hat{\phi}_i^-$. Putting our intermediate results together yields:

$$\begin{aligned}
&\mathbb{V}\left[\hat{\phi}_i \middle| \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}, \left(m_{i,\ell}^-\right)_{\ell\in\mathcal{L}^-}\right] \\
&= \mathbb{V}\left[\hat{\phi}_i^+ \mid \left(m_{i,\ell}^+\right)_{\ell\in\mathcal{L}^+}\right] + \mathbb{V}\left[\hat{\phi}_i^- \mid \left(m_{i,\ell}^-\right)_{\ell\in\mathcal{L}^-}\right] \\
&= \frac{1}{n^2}\sum_{\ell=1}^{n-3}\frac{{\sigma_{i,\ell}^+}^2}{m_{i,\ell}^+} + \frac{1}{n^2}\sum_{\ell=2}^{n-2}\frac{{\sigma_{i,\ell}^-}^2}{m_{i,\ell}^-} \\
&= \frac{1}{n^2}\sum_{\ell=1}^{n-3}\frac{{\sigma_{i,\ell}^+}^2}{m_{i,\ell}^+} + \frac{{\sigma_{i,\ell+1}^-}^2}{m_{i,\ell+1}^-}.
\end{aligned}$$

□

Lemma 13. *For any $i \in \mathcal{N}$ holds*

$$\mathbb{E}\left[\frac{1}{m_{i,\ell}^+}\right] \leq \frac{2n\log n}{\bar{T}} \text{ for all } \ell \in \mathcal{L}^+ \quad \text{and} \quad \mathbb{E}\left[\frac{1}{m_{i,\ell}^-}\right] \leq \frac{2n\log n}{\bar{T}} \text{ for all } \ell \in \mathcal{L}^-.$$

Proof. We prove the result only for $m_{i,\ell}^+$ since the proof for $m_{i,\ell}^-$ is analogous. By combining Equation (3.4) in Chao and Strawderman [1972]:

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1-(1-p)^{m+1}}{(m+1)p} \leq \frac{1}{mp} = \frac{1}{\mathbb{E}[X]},$$

for any binomially distributed random variable $X \sim Bin(m,p)$ with Lemma 10 and Lemma 11, we obtain:

$$\mathbb{E}\left[\frac{1}{m_{i,\ell}^+}\right] = \mathbb{E}\left[\frac{1}{1+\bar{m}_{i,\ell}^+}\right] \leq \frac{1}{\mathbb{E}\left[\bar{m}_{i,\ell}^+\right]} \leq \frac{2n\log n}{\bar{T}}.$$

□

Theorem 6 *For $\tilde{P}$ as chosen above the variance of any player's Shapley value estimate $\hat{\phi}_i$ is bounded by*

$$\mathbb{V}\left[\hat{\phi}_i\right] \leq \frac{2\log n}{n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^+}^2 + {\sigma_{i,\ell+1}^-}^2.$$

Proof. The combination of Lemma 12 and Lemma 13 yields:

$$\begin{aligned}
\mathbb{V}\left[\hat{\phi}_i\right] &= \mathbb{E}\left[\mathbb{V}\left[\hat{\phi}_i \middle| \left(m_{i,\ell}^{+}\right)_{\ell\in\mathcal{L}^+}, \left(m_{i,\ell}^{-}\right)_{\ell\in\mathcal{L}^-}\right]\right] \\
&\leq \mathbb{E}\left[\frac{1}{n^2}\sum_{\ell=1}^{n-3}\frac{{\sigma_{i,\ell}^{+}}^2}{m_{i,\ell}^{+}} + \frac{{\sigma_{i,\ell+1}^{-}}^2}{m_{i,\ell+1}^{-}}\right] \\
&= \frac{1}{n^2}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 \cdot \mathbb{E}\left[\frac{1}{m_{i,\ell}^{+}}\right] + {\sigma_{i,\ell+1}^{-}}^2 \cdot \mathbb{E}\left[\frac{1}{m_{i,\ell+1}^{-}}\right] \\
&\leq \frac{2\log n}{n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 + {\sigma_{i,\ell+1}^{-}}^2.
\end{aligned}$$

□

Corollary 2 *For $\tilde{P}$ as chosen above the MSE of any player's Shapley value estimate $\hat{\phi}_i$ is bounded by*

$$\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] \leq \frac{2\log n}{n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 + {\sigma_{i,\ell+1}^{-}}^2.$$

Proof. Using the bias-variance decomposition, the unbiasedness of $\hat{\phi}_i$ shown in Theorem 5, and the bound on the variance from Theorem 6 we obtain that:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] &= \left(\mathbb{E}\left[\hat{\phi}_i\right] - \phi_i\right)^2 + \mathbb{V}\left[\hat{\phi}_i\right] \\
&\leq \frac{2\log n}{n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 + {\sigma_{i,\ell+1}^{-}}^2.
\end{aligned}$$

□

D.4 Probabilistic Bounds

Theorem 7 *Fix any player $i \in \mathcal{N}$ and $\varepsilon > 0$. For $\tilde{P}$ as above the probability that the estimate $\hat{\phi}_i$ deviates from ϕ_i by a margin of ε or greater is bounded by*

$$\mathbb{P}\left(|\hat{\phi}_i - \phi| \geq \varepsilon\right) \leq \frac{2\log n}{\varepsilon^2 n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 + {\sigma_{i,\ell+1}^{-}}^2.$$

Proof. The bound on the variance of $\hat{\phi}_i$ in Theorem 6 allows us to apply Chebyshev's inequality:

$$\mathbb{P}\left(|\hat{\phi}_i - \phi| \geq \varepsilon\right) \leq \frac{\mathbb{V}\left[\hat{\phi}_i\right]}{\varepsilon^2} \leq \frac{2\log n}{\varepsilon^2 n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 + {\sigma_{i,\ell+1}^{-}}^2.$$

□

Corollary 4. *Fix any player $i \in \mathcal{N}$ and $\delta \in (0, 1]$. The estimate $\hat{\phi}_i$ deviates from ϕ_i by a margin of ε or greater with probability not greater than δ, i.e.,*

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq \delta \quad \textit{for} \quad \varepsilon = \sqrt{\frac{2\log n}{\delta n\bar{T}}\sum_{\ell=1}^{n-3}{\sigma_{i,\ell}^{+}}^2 + {\sigma_{i,\ell+1}^{-}}^2}.$$

Lemma 14. *For any $i \in \mathcal{N}$ and fixed $\varepsilon > 0$ holds:*

- $\mathbb{P}(|\hat{\phi}_{i,\ell}^{+} - \phi_{i,\ell}^{+}| \geq \varepsilon \mid m_{i,\ell}^{+}) \leq 2\exp\left(-\frac{2m_{i,\ell}^{+}\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)$ *for all* $\ell \in \mathcal{L}^+$

- $\mathbb{P}(|\hat{\phi}_{i,\ell}^- - \phi_{i,\ell}^-| \geq \varepsilon \mid m_{i,\ell}^-) \leq 2\exp\left(-\frac{2m_{i,\ell}^-\varepsilon^2}{{r_{i,\ell}^-}^2}\right)$ *for all* $\ell \in \mathcal{L}^-$

Proof. We prove the statement for $\hat{\phi}_{i,\ell}^+$ by making use of Hoeffding's inequality in combination with the unbiasedness of the strata estimates shown in Lemma 9. The proof for $\hat{\phi}_{i,\ell}^-$ is analogous.

$$
\begin{aligned}
&\mathbb{P}(|\hat{\phi}_{i,\ell}^+ - \phi_{i,\ell}^+| \geq \varepsilon \mid m_{i,\ell}^+) \\
&= \mathbb{P}\left(|\hat{\phi}_{i,\ell}^+ - \mathbb{E}[\hat{\phi}_{i,\ell}^+]| \geq \varepsilon \mid m_{i,\ell}^+\right) \\
&= \mathbb{P}\left(\left|\sum_{k=1}^{m_{i,\ell}^+} \nu(A_{i,\ell,k}^+) - \mathbb{E}\left[\sum_{k=1}^{m_{i,\ell}^+} \nu(A_{i,\ell,k}^+)\right]\right| \geq m_{i,\ell}^+\varepsilon \mid m_{i,\ell}^+\right) \\
&\leq 2\exp\left(-\frac{2m_{i,\ell}^+\varepsilon^2}{{r_{i,\ell}^+}^2}\right).
\end{aligned}
$$

□

Lemma 15. *For any* $i \in \mathcal{N}$ *and fixed* $\varepsilon > 0$ *holds:*

- $\mathbb{P}\left(|\hat{\phi}_{i,\ell}^+ - \phi_{i,\ell}^+| \geq \varepsilon\right) \leq \exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^+}^2}\right)^{\left\lfloor\frac{\bar{T}}{4n\log n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_{i,\ell}^+}^2}\right)-1}$ *for all* $\ell \in \mathcal{L}^+$
- $\mathbb{P}\left(|\hat{\phi}_{i,\ell}^- - \phi_{i,\ell}^-| \geq \varepsilon\right) \leq \exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^-}^2}\right)^{\left\lfloor\frac{\bar{T}}{4n\log n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_{i,\ell}^-}^2}\right)-1}$ *for all* $\ell \in \mathcal{L}^-$

Proof. We prove the statement for $\hat{\phi}_{i,\ell}^+$. The proof for $\hat{\phi}_{i,\ell}^-$ is analogous. To begin with, we derive with the help of Hoeffding's inequality for binomial distributions a bound for the probability of $\bar{m}_{i,\ell}^+$ not exceeding $\frac{\bar{T}}{4n\log n}$:

$$
\begin{aligned}
&\mathbb{P}\left(\bar{m}_{i,\ell}^+ \leq \frac{\bar{T}}{4n\log n}\right) \\
&\leq \mathbb{P}\left(\mathbb{E}\left[\bar{m}_{i,\ell}^+\right] - \bar{m}_{i,\ell}^+ \geq \mathbb{E}\left[\bar{m}_{i,\ell}^+\right] - \frac{\bar{T}}{4n\log n}\right) \\
&\leq \exp\left(-\frac{2\left(\mathbb{E}\left[\bar{m}_{i,\ell}^+\right] - \frac{\bar{T}}{4n\log n}\right)^2}{\bar{T}}\right) \\
&\leq \exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right),
\end{aligned}
$$

where we used the lower bound on $\mathbb{E}\left[\bar{m}_{i,\ell}^{+}\right]$ shown in Lemma 11. Next, we derive with the help of Lemma 14 a statement of technical nature to be used later:

$$\begin{aligned}
&\sum_{m=\lfloor \frac{\bar{T}}{4n\log n}\rfloor+1}^{\bar{T}} \mathbb{P}\left(|\hat{\phi}_{i,\ell}^{+}-\phi_{i,\ell}^{+}|\geq\varepsilon \mid m_{i,\ell}^{+}=m\right)\\
&\leq 2\sum_{m=\lfloor \frac{\bar{T}}{4n\log n}\rfloor+1}^{\bar{T}} \exp\left(-\frac{2m\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)\\
&= 2\sum_{m=0}^{\bar{T}} \exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)^m - 2\sum_{m=0}^{\lfloor \frac{\bar{T}}{4n\log n}\rfloor} \exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)^m\\
&= 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)^{\lfloor \frac{\bar{T}}{4n\log n}\rfloor} - \exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)^{\bar{T}}}{\exp\left(\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)-1}\\
&\leq 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)^{\lfloor \frac{\bar{T}}{4n\log n}\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)-1}.
\end{aligned}$$

At last, putting both findings together, we derive our claim:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\phi}_{i,\ell}^{+}-\phi_{i,\ell}^{+}|\geq\varepsilon\right)\\
&= \sum_{m=1}^{\bar{T}} \mathbb{P}\left(|\hat{\phi}_{i,\ell}^{+}-\phi_{i,\ell}^{+}|\geq\varepsilon \mid m_{i,\ell}^{+}=m\right)\cdot\mathbb{P}\left(m_{i,\ell}^{+}=m\right)\\
&= \sum_{m=1}^{\lfloor \frac{\bar{T}}{4n\log n}\rfloor} \mathbb{P}\left(|\hat{\phi}_{i,\ell}^{+}-\phi_{i,\ell}^{+}|\geq\varepsilon \mid m_{i,\ell}^{+}=m\right)\cdot\mathbb{P}\left(m_{i,\ell}^{+}=m\right)\\
&\quad + \sum_{m=\lfloor \frac{\bar{T}}{4n\log n}\rfloor+1}^{\bar{T}} \mathbb{P}\left(|\hat{\phi}_{i,\ell}^{+}-\phi_{i,\ell}^{+}|\geq\varepsilon \mid m_{i,\ell}^{+}=m\right)\cdot\mathbb{P}\left(m_{i,\ell}^{+}=m\right)\\
&\leq \mathbb{P}\left(\bar{m}_{i,\ell}^{+}\leq\left\lfloor\frac{\bar{T}}{4n\log n}\right\rfloor\right) + \sum_{m=\lfloor \frac{\bar{T}}{4n\log n}\rfloor+1}^{\bar{T}} \mathbb{P}\left(|\hat{\phi}_{i,\ell}^{+}-\phi_{i,\ell}^{+}|\geq\varepsilon \mid m_{i,\ell}^{+}=m\right)\\
&\leq \exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)^{\lfloor \frac{\bar{T}}{4n\log n}\rfloor}}{\exp\left(\frac{2\varepsilon^2}{{r_{i,\ell}^{+}}^2}\right)-1}.
\end{aligned}$$

□

Lemma 16. *For any $i\in\mathcal{N}$ and fixed $\varepsilon>0$ the probabilities that the estimates $\hat{\phi}_i^{+}$ and $\hat{\phi}_i^{-}$ deviate from ϕ_i^{+}, respectively ϕ_i^{-} are bounded by:*

- $\mathbb{P}\left(|\hat{\phi}_i^{+}-\phi_i^{+}|\geq\varepsilon\right)\leq(n-3)\left(\exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right)+2\frac{\exp\left(-\frac{2\varepsilon^2 n^2}{{R_i^{+}}^2}\right)^{\lfloor \frac{\bar{T}}{4n\log n}\rfloor}}{\exp\left(\frac{2\varepsilon^2 n^2}{{R_i^{+}}^2}\right)-1}\right)$,

- $\mathbb{P}\left(|\hat{\phi}_i^- - \phi_i^-| \geq \varepsilon\right) \leq (n-3)\left(\exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2 n^2}{R_i^{-2}}\right)^{\left\lfloor \frac{\bar{T}}{4n\log n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2 n^2}{R_i^{-2}}\right)-1}\right).$

Proof. We prove the statement for $\hat{\phi}_i^+$ using Lemma 15. The proof for $\hat{\phi}_i^-$ is analogous.

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{\ell=0}^{n-1} \hat{\phi}_{i,\ell}^+ - \phi_{i,\ell}^+\right| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(\frac{1}{n}\sum_{\ell=0}^{n-1} |\hat{\phi}_{i,\ell}^+ - \phi_{i,\ell}^+| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\sum_{\ell=1}^{n-3} |\hat{\phi}_{i,\ell}^+ - \phi_{i,\ell}^+| \geq \varepsilon n\right) \\
&\leq \sum_{\ell=1}^{n-3} \mathbb{P}\left(|\hat{\phi}_{i,\ell}^+ - \phi_{i,\ell}^+| \geq \frac{\varepsilon n r_{i,\ell}^+}{R^+}\right) \\
&\leq (n-3)\left(\exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2 n^2}{R^{+2}}\right)^{\left\lfloor \frac{\bar{T}}{4n\log n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2 n^2}{R^{+2}}\right)-1}\right)
\end{aligned}
$$

□

Theorem 8 *For any $i \in \mathcal{N}$ and fixed $\varepsilon > 0$ the probability that the estimate $\hat{\phi}_i$ deviates from ϕ_i by a margin of ε or greater is bounded by*

$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq 2(n-3)\left(\exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2 n^2}{(R_i^+ + R_i^-)^2}\right)^{\left\lfloor \frac{\bar{T}}{4n\log n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2 n^2}{(R_i^+ + R_i^-)^2}\right)-1}\right).$$

Proof. We apply Lemma 16 and obtain:

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \\
&= \mathbb{P}\left(|(\hat{\phi}_i^+ - \phi_i^+) + (\phi_i^- - \hat{\phi}_i^-)| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| + |\hat{\phi}_i^- - \phi_i^-| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(|\hat{\phi}_i^+ - \phi_i^+| \geq \frac{\varepsilon R_i^+}{R_i^+ + R_i^-}\right) + \mathbb{P}\left(|\hat{\phi}_i^- - \phi_i^-| \geq \frac{\varepsilon R_i^-}{R_i^+ + R_i^-}\right) \\
&\leq 2(n-3)\left(\exp\left(-\frac{\bar{T}}{8n^2(\log n)^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2 n^2}{(R_i^+ + R_i^-)^2}\right)^{\left\lfloor \frac{\bar{T}}{4n\log n}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2 n^2}{(R_i^+ + R_i^-)^2}\right)-1}\right).
\end{aligned}
$$

□

E Cooperative Games

E.1 Synthetic games

We provide formal definitions of the synthetic games and their Shapley values used in our empirical evaluation (see Section 6), and describe the process of how we randomly generated some of these.

E.1.1 Shoe Game

The number of players n in the Shoe game has to be even. The player set consist of two halves A and B of equal size, i.e., $\mathcal{N} = A \cup B$ with $A \cap B = \emptyset$ and $|A| = |B| = \frac{n}{2}$. The value function is given by $\nu(S) = \min\{|S \cap A|, |S \cap B|\}$. All players share the same Shapley value of $\phi_i = \frac{1}{2}$.

E.1.2 Airport Game

The Airport game entails $n = 100$ players. Each player i has an assigned weight c_i. The value function is the maximum of all weights contained in the coalition, i.e., $\nu(S) = \max_{i \in S} c_i$. The weights and resulting Shapley values are:

$$c_i = \begin{cases} 1 & \text{if } i \in \{1, \dots, 8\} \\ 2 & \text{if } i \in \{9, \dots, 20\} \\ 3 & \text{if } i \in \{21, \dots, 26\} \\ 4 & \text{if } i \in \{27, \dots, 40\} \\ 5 & \text{if } i \in \{41, \dots, 48\} \\ 6 & \text{if } i \in \{49, \dots, 57\} \\ 7 & \text{if } i \in \{58, \dots, 70\} \\ 8 & \text{if } i \in \{71, \dots, 80\} \\ 9 & \text{if } i \in \{81, \dots, 90\} \\ 10 & \text{if } i \in \{91, \dots, 100\} \end{cases}, \qquad \phi_i = \begin{cases} 0.01 & \text{if } i \in \{1, \dots, 8\} \\ 0.020869565 & \text{if } i \in \{9, \dots, 20\} \\ 0.033369565 & \text{if } i \in \{21, \dots, 26\} \\ 0.046883079 & \text{if } i \in \{27, \dots, 40\} \\ 0.063549745 & \text{if } i \in \{41, \dots, 48\} \\ 0.082780515 & \text{if } i \in \{49, \dots, 57\} \\ 0.106036329 & \text{if } i \in \{58, \dots, 70\} \\ 0.139369662 & \text{if } i \in \{71, \dots, 80\} \\ 0.189369662 & \text{if } i \in \{81, \dots, 90\} \\ 0.289369662 & \text{if } i \in \{91, \dots, 100\} \end{cases}.$$

E.1.3 SOUG Game

A Sum of unanimity games (SOUG) is specified by M many sets $S_1, \dots, S_M \subseteq \mathcal{N}$ and weights $c_1, \dots, c_M \in \mathbb{R}$. The value functions is defined as $\nu(S) = \sum_{m=1}^{M} c_m \cdot \mathbb{I}_{S_m \subseteq S}$ leading to Shapley values $\phi_i = \sum_{m=1}^{M} \frac{c_m}{|S_m|} \cdot \mathbb{I}_{i \in S_m}$, which can be computed in polynomial time if knowledge of sets and coefficients is provided. We generate SOUG games with $M = 50$ randomly by selecting for each S_m to be drawn a size uniformly at random between 1 and n, and then draw the set S_m with that size uniformly. We draw the coefficients uniformly at random from $[0, 1]$.

E.2 Explainability games

In the following, we describe the three explainability games introduced in Section 6; namely, the NLP sentiment analysis game (see Section E.2.1), the image classifier game (see Section E.2.2), and the adult classification (see Section E.2.3), and explain the value function of each resulting cooperative game. Since there exists no efficient closed-form solution, we compute the Shapley values exhaustively (via brute force) in order to allow the tracking of the approximation error of the different algorithms. Due to constraints in computational power this limits us to $n = 14$ players per game, which necessitates $2^{14} = 16\,384$ model evaluation to exhaustively traverse the powerset of all coalitions. Note that after a budget of $T = 2^{14}$ both KernelSHAP and Stratified SVARM$^+$ have an approximation error of zero because both have observed all coalition values and thus have seen the cooperative game in its entirety.

E.2.1 NLP sentiment analysis

The NLP sentiment game describes an explainability scenario for local feature importance of a sentiment classification model. The sentiment classifier[3] is a fine-tuned version of the DistilBERT transformer architecture Sanh et al. [2019]. The model was fine-tuned on the *IMDB* dataset Maas et al. [2011]. The model expects a natural language sentence as input, transforms the sentence into a tokenized form and predicts a sentiment score ranging from $[-1, 1]$. We randomly select sentences from the *IMDB* dataset, that contain no more than 14 tokens. For a sentence the local explainability

[3] `https://huggingface.co/dhlee347/distilbert-imdb`.

game consists of presenting the model a coalition of players (tokens) and observing the predicted sentiment as the value of a coalition. Absent players are removed in the tokenized representation (i.e. tokens are removed).

E.2.2 Image classifier

The image classifier game is similar to the NLP sentiment analysis game (Section E.2.1) a local explanation scenario. For this we explain the output of an image classifier given random images from ImageNet Deng et al. [2009]. The model to be explained is a ResNet18[4] He et al. [2016] trained on ImageNet Deng et al. [2009]. To restrict the number of players, we apply SLIC Achanta et al. [2012] to summarize individual pixels with 14 super-pixels. The super-pixels then make up the players in the image classification game. A coalition of players, thus, consists of the corresponding super-pixels. The super-pixels of absent players are removed via mean-imputation by setting all their pixels to grey. The worth of a coalition is determined by the output of the model (using only the present super-pixels given by the coalition) for the class of the original prediction which was made with all pixels being present.

E.2.3 Adult classification

Similar to the preceding two games, the adult classification game is also a local explanation scenario. We train a gradient-boosted tree classifier (sklearn) on the adult dataset Becker and Kohavi [1996] to classify whether an adult has an income below or above $50\,000$. Each game is based on a randomly chosen datapoint for which the players correspond to features. A coalition is formed by removing the absent feature values of the selected datapoint via mean imputation. The worth of a coalition is the predicted class probability of the true income class given the manipulated datapoint after mean imputation of absent features.

[4] `https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html`.

F Further Empirical Results

The plots shown in Figure 3 hardly visualize the performance differences between the algorithms with low MSE values. Thus, we present our findings in Figure 4 to Figure 9 in higher resolution.

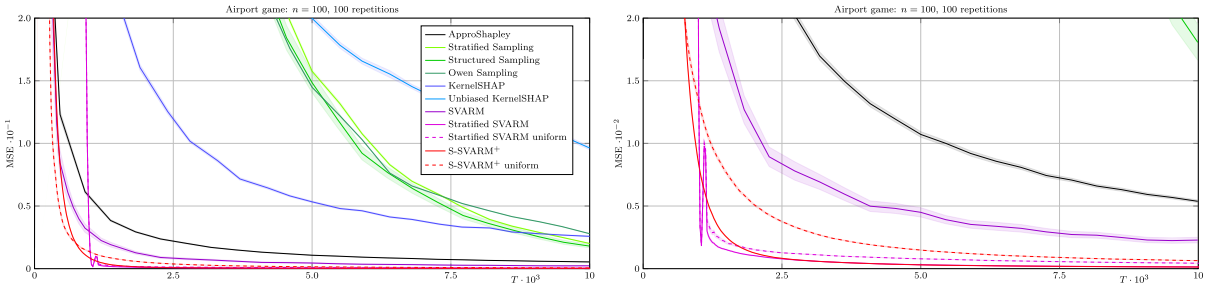


Figure 4: Airport game with 100 players: Averaged MSE over 100 repetitions in dependence of fixed budget T, shaded bands showing standard errors.

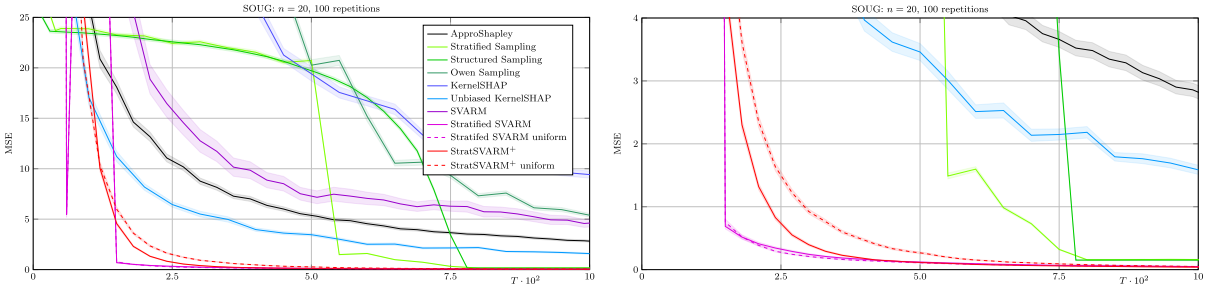


Figure 5: SOUG game with 20 players: Averaged MSE over 100 repetitions in dependence of fixed budget T, shaded bands showing standard errors.

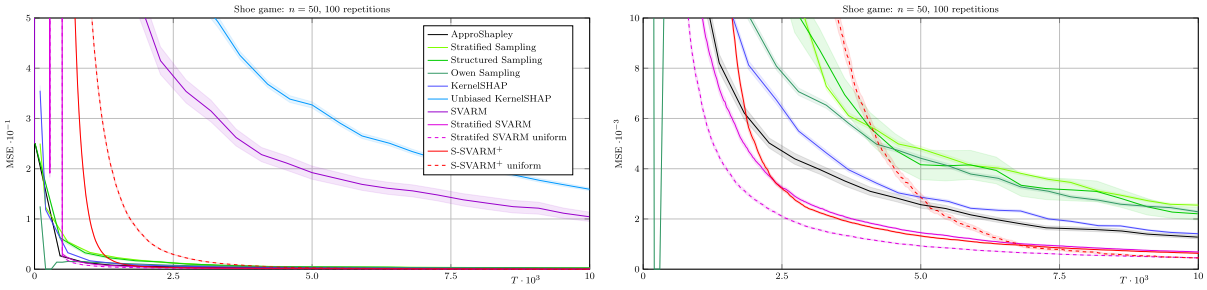


Figure 6: Shoe game with 50 players: Averaged MSE over 100 repetitions in dependence of fixed budget T, shaded bands showing standard errors.

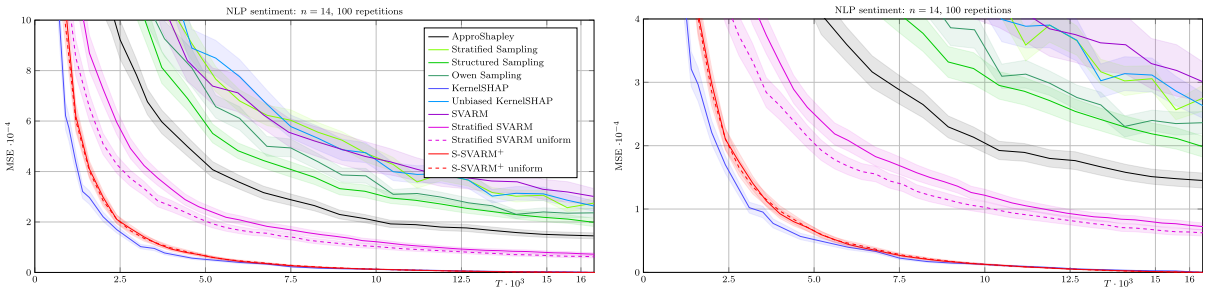


Figure 7: NLP game with 14 players: Averaged MSE over 100 repetitions in dependence of fixed budget T, shaded bands showing standard errors.

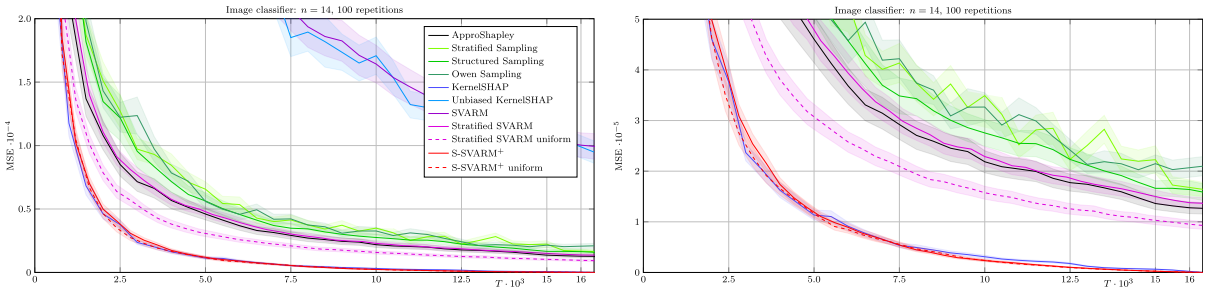


Figure 8: Image classifier game with 14 players: Averaged MSE over 100 repetitions in dependence of fixed budget T, shaded bands showing standard errors.

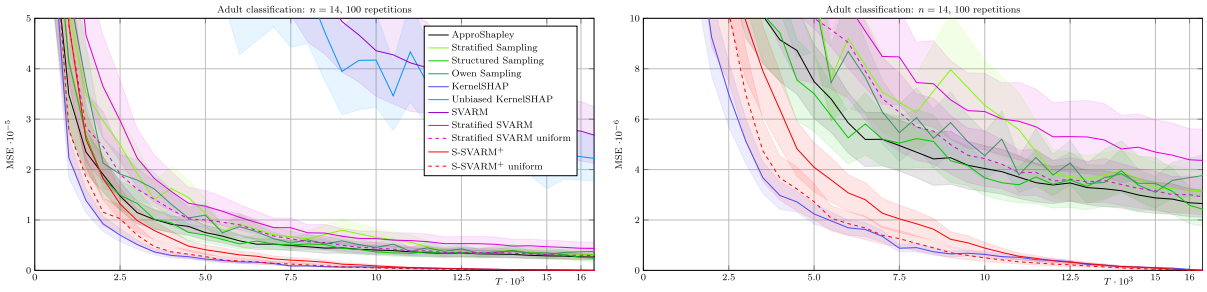


Figure 9: Adult classification with 14 players: Averaged MSE over 100 repetitions in dependence of fixed budget T, shaded bands showing standard errors.

B

Appendix to Shapley Value Approximation Based on k-Additive Games

A. Theoretical Analysis

In the following we prove Theorem 4.2 by solving the k-additive optimization problem with weights $w_A = \binom{n-2}{|A|-1}^{-1}$ analytically and showing that the solution contains the Shapley value. We introduce some simplifying notation:

- $\tilde{\mathcal{P}} := \mathcal{P}(N) \setminus \{\emptyset, N\}$
- $I_0 := I^k(\emptyset)$, $I_i := I^k(\{i\})$ for all $i \in N$, $I_{i,j} := I^k(\{i,j\})$ for all $\{i,j\} \subseteq N$, $I_{i,j,\ell} := I^k(\{i,j,\ell\})$
- Weight w_a for any A with $|A| = a$
- $\gamma_{A,B} := \gamma^{|B|}_{|A \cap B|}$ for all $A, B \subseteq N$
- $\beta_{i,A} = \begin{cases} 1 & \text{if } i \notin A \\ -1 & \text{if } i \in A \end{cases}$ for all $i \in N$ and $A \in \mathcal{P}(N) \setminus \{\emptyset, N\}$
- $\beta_{i,j,A} = \begin{cases} 2 & \text{if } |\{i,j\} \cap A| = 1 \\ -1 & \text{otherwise} \end{cases}$ for all $\{i,j\} \subseteq N$ and $A \in \mathcal{P}(N) \setminus \{\emptyset, N\}$
- $\beta_{i,j,\ell,A} = \begin{cases} -1 & \text{if } |\{i,j,\ell\} \cap A| = 1 \\ 1 & \text{if } |\{i,j,\ell\} \cap A| = 2 \\ 0 & \text{otherwise} \end{cases}$ for all $\{i,j,\ell\} \subseteq N$ and $A \in \mathcal{P}(N) \setminus \{\emptyset, N\}$

Our proof is preceded by an observation that we shall utilize later:

Lemma A.1. *For any set of players N, player $i \in N$ and for the cases $\ell = 2$ and $\ell = 3$ the following equality holds:*

$$\sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \sum_{\substack{B \subseteq N \\ |B| = \ell}} \gamma_{A,B} I_B = \frac{1}{n} \sum_{j \in N} \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{j,A} \sum_{\substack{B \subseteq N \\ |B| = \ell}} \gamma_{A,B} I_B \,.$$

Proof:
We show the statement for both cases separately and start with $\ell = 2$. For interactions I_{j_1,j_2} that contain i we derive after inserting the weights $w_a = \binom{n-2}{a-1}^{-1}$:

$$\begin{aligned}
& \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \sum_{\substack{\{j_1,j_2\} \subseteq N \\ i \in \{j_1,j_2\}}} \beta_{j_1,j_2,A} I_{j_1,j_2} \\
= & \sum_{j_1 \in N \setminus \{i\}} I_{i,j_1} \sum_{A \in \tilde{\mathcal{P}}} \beta_{i,A} \beta_{j_1,i,A} w_A \\
= & \sum_{j_1 \in N \setminus \{i\}} I_{i,j_1} \left(\sum_{\substack{A \in \tilde{\mathcal{P}} \\ i,j_1 \in A}} w_A - \sum_{\substack{A \in \tilde{\mathcal{P}} \\ i,j_1 \notin A}} w_A - 2 \sum_{\substack{A \in \tilde{\mathcal{P}} \\ i \in A, j_1 \notin A}} w_A + 2 \sum_{\substack{A \in \tilde{\mathcal{P}} \\ i \notin A, j_1 \in A}} w_A \right) \\
= & \sum_{j_1 \in N \setminus \{i\}} I_{i,j_1} \left(\sum_{a=2}^{n-1} \sum_{\substack{A \in \tilde{\mathcal{P}}, |A| = a \\ i,j_1 \in A}} w_a - \sum_{a=1}^{n-2} \sum_{\substack{A \in \tilde{\mathcal{P}}, |A| = a \\ i,j_1 \notin A}} w_a - 2 \sum_{a=1}^{n-1} \sum_{\substack{A \in \tilde{\mathcal{P}}, |A| = a \\ i \in A, j_1 \notin A}} w_a + 2 \sum_{a=1}^{n-1} \sum_{\substack{A \in \tilde{\mathcal{P}}, |A| = a \\ i \notin A, j_1 \in A}} w_a \right) \qquad (10) \\
= & \sum_{j_1 \in N \setminus \{i\}} I_{i,j_1} \left(\sum_{a=2}^{n-1} \tbinom{n-2}{a-2} w_a - \sum_{a=1}^{n-2} \tbinom{n-2}{a} w_a - 2 \sum_{a=1}^{n-1} \tbinom{n-2}{a-1} w_a + 2 \sum_{a=1}^{n-1} \tbinom{n-2}{a-1} w_a \right) \\
= & \sum_{j_1 \in N \setminus \{i\}} I_{i,j_1} \sum_{a=1}^{n-1} \left(\tbinom{n-2}{a-2} - \tbinom{n-2}{a} \right) w_a \\
= & \sum_{j_1 \in N, j_1 \neq i} I_{j_1,i} \sum_{a=1}^{n-1} \tfrac{a-1}{n-a} - \tfrac{n-a-1}{a} \\
= & \; 0
\end{aligned}$$

And for all other interactions I_{j_1,j_2} not containing i we derive:

$$
\begin{aligned}
& \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} \beta_{j_1,j_2,A} I_{j_1,j_2} \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{j_1,j_2}\Bigg(\sum_{\substack{A\in\tilde{\mathcal{P}}\\ i,j_1,j_2\in A}} w_A + \sum_{\substack{A\in\tilde{\mathcal{P}}\\ i\in A, j_1,j_2\notin A}} w_A - 2 \sum_{\substack{A\in\tilde{\mathcal{P}}\\ i,j_1\in A, j_2\notin A}} w_A - 2 \sum_{\substack{A\in\tilde{\mathcal{P}}\\ i,j_2\in A, j_1\notin A}} w_A \\
& - \sum_{\substack{A\in\tilde{\mathcal{P}}\\ i\notin A, j_1,j_2\in A}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}}\\ i,j_1,j_2\notin A}} w_A + 2 \sum_{\substack{A\in\tilde{\mathcal{P}}\\ j_1\in A, i,j_2\notin A}} w_A + 2 \sum_{\substack{A\in\tilde{\mathcal{P}}\\ j_2\in A, i,j_1\notin A}} w_A \Bigg) \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{j_1,j_2}\Bigg(\sum_{a=3}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i,j_1,j_2\in A}} w_a + \sum_{a=1}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i\in A, j_1,j_2\notin A}} w_a - 2\sum_{a=2}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i,j_1\in A, j_2\notin A}} w_a - 2\sum_{a=2}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i,j_2\in A, j_1\notin A}} w_a \\
& - \sum_{a=2}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i\notin A, j_1,j_2\in A}} w_a - \sum_{a=1}^{n-3} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i,j_1,j_2\notin A}} w_a + 2\sum_{a=1}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ j_1\in A, i,j_2\notin A}} w_a + 2\sum_{a=1}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ j_2\in A, i,j_1\notin A}} w_a \Bigg) \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{j_1,j_2}\Bigg(\sum_{a=3}^{n-1} \tbinom{n-3}{a-3} w_a + \sum_{a=1}^{n-2} \tbinom{n-3}{a-1} w_a - 2\sum_{a=2}^{n-1} \tbinom{n-3}{a-2} w_a - 2\sum_{a=2}^{n-1} \tbinom{n-3}{a-2} w_a \\
& - \sum_{a=2}^{n-1} \tbinom{n-3}{a-2} w_a - \sum_{a=1}^{n-3} \tbinom{n-3}{a} w_a + 2\sum_{a=1}^{n-2} \tbinom{n-3}{a-1} w_a + 2\sum_{a=1}^{n-2} \tbinom{n-3}{a-1} w_a \Bigg) \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{j_1,j_2} \sum_{a=1}^{n-1} \left(\tbinom{n-3}{a-3} - 5\tbinom{n-3}{a-2} + 5\tbinom{n-3}{a-1} - \tbinom{n-3}{a} \right) w_a \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{j_1,j_2} \sum_{a=1}^{n-1} \tfrac{(a-1)(a-2)}{(n-2)(n-a)} - 5\tfrac{a-1}{n-2} + 5\tfrac{n-a-1}{n-2} - \tfrac{(n-a-1)(n-a-2)}{(n-2)a} \\
= & \; 0
\end{aligned}
\tag{11}
$$

Adding Equation (10) and (11) yields:

$$
\begin{aligned}
& \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\substack{B\subseteq N\\ |B|=2}} \gamma_{A,B} I_B \\
= & -\tfrac{1}{6} \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\{j_1,j_2\}\subseteq N} \beta_{j_1,j_2,A} I_{j_1,j_2} \\
= & -\tfrac{1}{6} \left(\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\substack{\{j_1,j_2\}\subseteq N\\ i\in\{j_1,j_2\}}} \beta_{j_1,j_2,A} I_{j_1,j_2} + \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} \beta_{j_1,j_2,A} I_{j_1,j_2} \right) \\
= & \; 0
\end{aligned}
$$

Consequently, we also have

$$
\tfrac{1}{n} \sum_{j\in N} \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{j,A} \sum_{\substack{B\subseteq N\\ |B|=2}} \gamma_{A,B} I_B = 0\,.
$$

Continuing with $\ell = 3$, for interactions I_{j_1,j_2,j_3} that contain i we derive after inserting the weights $w_a = \binom{n-2}{a-1}^{-1}$:

$$
\begin{aligned}
& \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\substack{\{j_1,j_2,j_3\}\subseteq N\\ i\in\{j_1,j_2,j_3\}}} \beta_{j_1,j_2,j_3,A} I_{j_1,j_2,j_3} \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2} \sum_{A\in\tilde{P}} \beta_{i,A}\beta_{i,j_1,j_2} w_A \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2} \left(\sum_{\substack{A\in\tilde{\mathcal{P}},i\in A\\ |\{j_1,j_2\}\cap A|=0}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}},i\in A\\ |\{j_1,j_2\}\cap A|=1}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}},i\notin A\\ |\{j_1,j_2\}\cap A|=1}} w_A + \sum_{\substack{A\in\tilde{\mathcal{P}},i\notin A\\ |\{j_1,j_2\}\cap A|=2}} w_A \right) \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2} \left(\sum_{a=1}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i\in A,|\{j_1,j_2\}\cap A|=0}} w_a - \sum_{a=2}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i\in A,|\{j_1,j_2\}\cap A|=1}} w_a \right. \\
& \left. - \sum_{a=1}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i\notin A,|\{j_1,j_2\}\cap A|=1}} w_a + \sum_{a=2}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\ i\notin A,|\{j_1,j_2\}\cap A|=2}} w_a \right) \\
= & \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2} \left(\sum_{a=1}^{n-2} \tbinom{n-3}{a-1} w_a - 2\sum_{a=2}^{n-1} \tbinom{n-3}{a-2} w_a - 2\sum_{a=1}^{n-2} \tbinom{n-3}{a-1} w_a + \sum_{a=2}^{n-1} \tbinom{n-3}{a-2} w_a \right) \\
= & - \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2} \sum_{a=1}^{n-1} \tbinom{n-2}{a-1} w_a \\
= & -(n-1) \sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2}
\end{aligned}
\tag{12}
$$

And for all other interactions I_{j_1,j_2,j_3} not containing i we derive:

$$
\begin{aligned}
& \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A} \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} \beta_{j_1,j_2,j_3,A} I_{j_1,j_2,j_3} \\
= & \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3} \sum_{A\in\tilde{\mathcal{P}}} \beta_{i,A}\beta_{j_1,j_2,j_3,A} w_A \\
= & \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3} \left(\sum_{\substack{A\in\tilde{\mathcal{P}},i\in A\\ |\{j_1,j_2,j_3\}\cap A|=1}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}},i\in A\\ |\{j_1,j_2,j_3\}\cap A|=2}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}},i\notin A\\ |\{j_1,j_2,j_3\}\cap A|=1}} w_A + \sum_{\substack{A\in\tilde{\mathcal{P}},i\notin A\\ |\{j_1,j_2,j_3\}\cap A|=2}} w_A \right) \\
= & \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3} \left(\sum_{a=2}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a,i\in A\\ |\{j_1,j_2,j_3\}\cap A|=1}} w_a - \sum_{a=3}^{n-1} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a,i\in A\\ |\{j_1,j_2,j_3\}\cap A|=2}} w_a \right. \\
& \left. - \sum_{a=1}^{n-3} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a,i\notin A\\ |\{j_1,j_2,j_3\}\cap A|=1}} w_a + \sum_{a=2}^{n-2} \sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a,i\notin A\\ |\{j_1,j_2,j_3\}\cap A|=2}} w_a \right) \\
= & 3 \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3} \left(\sum_{a=2}^{n-2} \tbinom{n-4}{a-2} w_a - \sum_{a=3}^{n-1} \tbinom{n-4}{a-3} w_a - \sum_{a=1}^{n-3} \tbinom{n-4}{a-1} w_a + \sum_{a=2}^{n-2} \tbinom{n-4}{a-2} w_a \right) \\
= & 3 \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3} \sum_{a=1}^{n-1} \left(4\tbinom{n-4}{a-2} - \tbinom{n-2}{a-1} \right) w_a \\
= & 3 \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3} \sum_{a=1}^{n-1} 4\tfrac{(a-1)(n-a-1)}{(n-2)(n-3)} - 1 \\
= & -(n-1) \sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3}
\end{aligned}
\tag{13}
$$

Adding Equation (12) and (13) yields:

$$\begin{aligned}
&\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{\substack{B\subseteq N\\|B|=3}}\gamma_{A,B}I_B\\
=\;& -\tfrac{1}{6}\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{\{j_1,j_2,j_3\}\subseteq N}\beta_{j_1,j_2,j_3,A}I_{j_1,j_2,j_3}\\
=\;& -\tfrac{1}{6}\left(\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{\substack{\{j_1,j_2,j_3\}\subseteq N\\ i\in\{j_1,j_2,j_3\}}}\beta_{j_1,j_2,j_3,A}I_{j_1,j_2,j_3}+\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}}\beta_{j_1,j_2,j_3,A}I_{j_1,j_2,j_3}\right)\\
=\;& \tfrac{n-1}{6}\left(\sum_{\{j_1,j_2\}\subseteq N\setminus\{i\}} I_{i,j_1,j_2}+\sum_{\{j_1,j_2,j_3\}\subseteq N\setminus\{i\}} I_{j_1,j_2,j_3}\right)\\
=\;& \tfrac{n-1}{6}\sum_{\{j_1,j_2,j_3\}\subseteq N} I_{j_1,j_2,j_3}
\end{aligned}$$

Obviously, summing up the last term over all $j\in N$ and dividing it by n will not change it, which concludes the proof. □

Proof of Theorem 4.2:
The constraint to guarantee the efficiency axiom can be simplified, leading to the following optimization problem:

$$\begin{aligned}
\min_I\quad & \sum_{A\in\tilde{\mathcal{P}}} w_A\left(\nu(A)-\sum_{B\subseteq N,|B|\le k}\gamma_{A,B}I_B\right)^2\\
\text{s.t.}\quad & \nu(N)-\nu(\emptyset)=\sum_{i\in N} I_i
\end{aligned}$$

We apply the Lagrange method. The new objective to minimize is

$$\Lambda(I,\lambda)=\sum_{A\in\tilde{\mathcal{P}}} w_A\left(\nu(A)-\sum_{B\subseteq N,|B|\le k}\gamma_{A,B}I_B\right)^2+\lambda\left(\sum_{i\in N} I_i-\nu(N)+\nu(\emptyset)\right).$$

The partial derivatives of Λ must turn to zero for its solution. Hence we obtain the following equations:

$$\begin{aligned}
\tfrac{\partial\Lambda}{\partial I_0} &= -2\sum_{A\in\tilde{\mathcal{P}}} w_A\gamma_{A,\emptyset}\left(\nu(A)-\sum_{B\subseteq N,|B|\le k}\gamma_{A,B}I_B\right) &&\stackrel{!}{=}0\\
\tfrac{\partial\Lambda}{\partial I_i} &= -2\sum_{A\in\tilde{\mathcal{P}}} w_A\gamma_{A,\{i\}}\left(\nu(A)-\sum_{B\subseteq N,|B|\le k}\gamma_{A,B}I_B\right)+\lambda &&\stackrel{!}{=}0\ \text{ for all } i\in N\\
\tfrac{\partial\Lambda}{\partial I_S} &= -2\sum_{A\in\tilde{\mathcal{P}}} w_A\gamma_{A,S}\left(\nu(A)-\sum_{B\subseteq N,|B|\le k}\gamma_{A,B}I_B\right) &&\stackrel{!}{=}0\ \text{ for all } S\subseteq N \text{ with } |S|\in[2,k]\\
\tfrac{\partial\Lambda}{\partial\lambda} &= \sum_{i\in N} I_i-\nu(N)+\nu(\emptyset) &&\stackrel{!}{=}0
\end{aligned}$$

From $\frac{\partial\Lambda}{\partial\lambda}$ we immediately extract

$$\sum_{i\in N} I_i=\nu(N)-\nu(\emptyset) \tag{14}$$

and thus also for any $i\in N$:

$$\sum_{j\in N\setminus\{i\}} I_j=\nu(N)-\nu(\emptyset)-I_i\,. \tag{15}$$

The derivative $\frac{\partial\Lambda}{\partial I_i}$ can be rearranged for any $i\in N$ to obtain an expression of summands grouped by the order of their contained interactions:

$$\begin{aligned}
\tfrac{\partial\Lambda}{\partial I_i} &= \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\left(\nu(A)-\sum_{B\subseteq N,|B|\le k}\gamma_{A,B}I_B\right)+\lambda\\
&= \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\nu(A)-\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}I_0+\tfrac{1}{2}\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{j\in N}\beta_{j,A}I_j-\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{\substack{B\subseteq N\\2\le|B|\le k}}\gamma_{A,B}I_B+\lambda
\end{aligned} \tag{16}$$

In the following, we derive expressions for multiple terms that are contained in Equation (16). First, we have for the sum containing the interaction of the empty set:

$$\begin{aligned}\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}I_0 &= I_0\left(\sum_{A\in\tilde{\mathcal{P}},i\notin A} w_A - \sum_{A\in\tilde{\mathcal{P}},i\in A} w_A\right)\\ &= I_0\left(\sum_{a=1}^{n-1}\sum_{\substack{A\in\tilde{\mathcal{P}},i\notin A\\|A|=a}} w_a - \sum_{a=1}^{n-1}\sum_{\substack{A\in\tilde{\mathcal{P}},i\in A\\|A|=a}} w_a\right)\\ &= I_0\sum_{a=1}^{n-1}\left(\tbinom{n-1}{a} - \tbinom{n-1}{a-1}\right)w_a\\ &= I_0\sum_{a=1}^{n-1}\tfrac{n-1}{a} - \tfrac{n-1}{n-a}\\ &= 0\end{aligned} \tag{17}$$

Next, we solve the sum that contains interactions of singletons, requiring two steps, we begin with

$$\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}^2 I_i = I_i\sum_{a=1}^{n-1}\tbinom{n}{a}w_a\,. \tag{18}$$

In the second step we analyze and utilize Equation (15) to obtain:

$$\begin{aligned}&\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{j\in N,j\neq i}\beta_{j,A}I_j\\ =\;& \sum_{j\in N,j\neq i} I_j\left(\sum_{\substack{A\in\tilde{\mathcal{P}}\\i,j\in A}} w_A + \sum_{\substack{A\in\tilde{\mathcal{P}}\\i,j\notin A}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}}\\i\in A,j\notin A}} w_A - \sum_{\substack{A\in\tilde{\mathcal{P}}\\i\notin A,j\in A}} w_A\right)\\ =\;& \sum_{j\in N,j\neq i} I_j\left(\sum_{a=2}^{n-1}\sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\i,j\in A}} w_a + \sum_{a=1}^{n-2}\sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\i,j\notin A}} w_a - \sum_{a=1}^{n-1}\sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\i\in A,j\notin A}} w_a - \sum_{a=1}^{n-1}\sum_{\substack{A\in\tilde{\mathcal{P}},|A|=a\\i\notin A,j\in A}} w_a\right)\\ =\;& \sum_{j\in N,j\neq i} I_j\left(\sum_{a=2}^{n-1}\tbinom{n-2}{a-2}w_a + \sum_{a=1}^{n-2}\tbinom{n-2}{a}w_a - 2\sum_{a=1}^{n-1}\tbinom{n-2}{a-1}w_a\right)\\ =\;& (\nu(N)-\nu(\emptyset)-I_i)\sum_{a=1}^{n-1}\left(\tbinom{n-2}{a-2} + \tbinom{n-2}{a} - 2\tbinom{n-2}{a-1}\right)w_a\end{aligned} \tag{19}$$

We combine Equation (18) with Equation (19), and apply the weights $w_a = \binom{n-2}{a-1}^{-1}$ and the identity $2H_{n-1} = \sum_{a=1}^{n-1}\frac{n}{a(n-a)}$, where $H_{n-1} = \sum_{a=1}^{n-1}\frac{1}{a}$ is the harmonic sum, to derive:

$$\begin{aligned}&\sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{j\in N}\beta_{j,A}I_j\\ =\;& \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\left(\beta_{i,A}I_i + \sum_{j\in N,j\neq i}\beta_{j,A}I_j\right)\\ =\;& \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}^2 I_i + \sum_{A\in\tilde{\mathcal{P}}} w_A\beta_{i,A}\sum_{j\in N,j\neq i}\beta_{j,A}I_j\\ =\;& I_i\sum_{a=1}^{n-1}\tbinom{n}{a}w_a + (\nu(N)-\nu(\emptyset)-I_i)\sum_{a=1}^{n-1}\left(\tbinom{n-2}{a-2} + \tbinom{n-2}{a} - 2\tbinom{n-2}{a-1}\right)w_a\\ =\;& I_i\sum_{a=1}^{n-1}\left(\tbinom{n}{a} - \tbinom{n-2}{a-2} - \tbinom{n-2}{a} + 2\tbinom{n-2}{a-1}\right)w_a + (\nu(N)-\nu(\emptyset))\sum_{a=1}^{n-1}\left(\tbinom{n-2}{a-2} + \tbinom{n-2}{a} - 2\tbinom{n-2}{a-1}\right)w_a\\ =\;& 4I_i\sum_{a=1}^{n-1}\tbinom{n-2}{a-1}w_a + (\nu(N)-\nu(\emptyset))\sum_{a=1}^{n-1}\left(\tbinom{n}{a} - 4\tbinom{n-2}{a-1}\right)w_a\\ =\;& 4(n-1)I_i + (\nu(N)-\nu(\emptyset))\sum_{a=1}^{n-1}\left(\tfrac{n(n-1)}{a(n-a)} - 4\right)\\ =\;& 4(n-1)I_i + 2(n-1)(H_{n-1}-2)\,(\nu(N)-\nu(\emptyset))\end{aligned} \tag{20}$$

Summing Equation (20) up over all $i \in N$ yields under usage of Equation (14):

$$\sum_{i \in N} \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \sum_{j \in N} \beta_{j,A} I_j = 2(n-1)(n(H_{n-1}-2)+2)\left(\nu(N)-\nu(\emptyset)\right) . \tag{21}$$

And as the final intermediate term we derive for the weighted coalition values summed up over all $i \in N$:

$$\sum_{i \in N} \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \nu(A) = \sum_{A \in \tilde{\mathcal{P}}} w_A \nu(A)\left(n - 2|A|\right) . \tag{22}$$

For any $i \in N$, after rearranging Equation (16) and plugging in Equation (17) and (20) we have

$$-\lambda = \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \nu(A) + 2(n-1) I_i + (n-1)(H_{n-1}-2)\left(\nu(N)-\nu(\emptyset)\right) - \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \sum_{\substack{B \subseteq N \\ 2 \leq |B| \leq k}} \gamma_{A,B} I_B . \tag{23}$$

We also obtain for $-\lambda$ by rearranging Equation (16), summing it up over all $i \in N$, dividing by n, and plugging in Equation (17), (21), and (22):

$$-\lambda = \tfrac{1}{n} \sum_{A \in \tilde{\mathcal{P}}} w_A \nu(A)\left(n - 2|A|\right) + \tfrac{n-1}{n}(n(H_{n-1}-2)+2)\left(\nu(N)-\nu(\emptyset)\right) - \tfrac{1}{n} \sum_{i \in N} \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \sum_{\substack{B \subseteq N \\ 2 \leq |B| \leq k}} \gamma_{A,B} I_B . \tag{24}$$

Finally, we conclude the proof by equating Equation (23) and (24). We utilize Lemma A.1 to cancel out the sums that contain interactions of order two and three such that the theorem holds true for the cases of $k = 2$ and $k = 3$. This step is does not apply for the special case of $k = 1$ since the last sums in Equation (23) and (24) vanish. We solve for I_i and derive:

$$\begin{aligned}
I_i = \;& \tfrac{1}{2n}(n(H_{n-1}-2)+2)\left(\nu(N)-\nu(\emptyset)\right) - \tfrac{H_{n-1}-2}{2}\left(\nu(N)-\nu(\emptyset)\right) \\
& + \tfrac{1}{2n(n-1)} \sum_{A \in \tilde{\mathcal{P}}} w_A \nu(A)\left(n-2|A|\right) - \tfrac{1}{2(n-1)} \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \nu(A) \\
& + \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{i,A} \sum_{\substack{B \subseteq N \\ 2 \leq |B| \leq k}} \gamma_{A,B} I_B - \tfrac{1}{n} \sum_{j \in N} \sum_{A \in \tilde{\mathcal{P}}} w_A \beta_{j,A} \sum_{\substack{B \subseteq N \\ 2 \leq |B| \leq k}} \gamma_{A,B} I_B \\
= \;& \tfrac{1}{n}\left(\nu(N)-\nu(\emptyset)\right) + \tfrac{1}{2n(n-1)} \sum_{A \in \tilde{\mathcal{P}} i \in A} w_A \nu(A)\left(n-2|A|\right) + \tfrac{1}{2(n-1)} \sum_{A \in \tilde{\mathcal{P}}, i \in A} w_A \nu(A) \\
& + \tfrac{1}{2n(n-1)} \sum_{A \in \tilde{\mathcal{P}}, i \notin A} w_A \nu(A)\left(n-2|A|\right) - \tfrac{1}{2(n-1)} \sum_{A \in \tilde{\mathcal{P}}, i \notin A} w_A \nu(A) \\
= \;& \tfrac{1}{n}\left(\nu(N)-\nu(\emptyset)\right) + \sum_{A \in \tilde{\mathcal{P}}, i \in A} \tfrac{n-|A|}{n(n-1)} \cdot w_A \nu(A) - \sum_{A \in \tilde{\mathcal{P}}, i \notin A} \tfrac{|A|}{n(n-1)} \cdot w_A \nu(A) \\
= \;& \tfrac{1}{n}\left(\nu(N)-\nu(\emptyset)\right) + \sum_{A \in \tilde{\mathcal{P}}, i \in A} \tfrac{1}{n \cdot \binom{n-1}{|A|-1}} \cdot \nu(A) - \sum_{A \in \tilde{\mathcal{P}}, i \notin A} \tfrac{1}{n \cdot \binom{n-1}{|A|}} \nu(A) \\
= \;& \sum_{A \subseteq N, i \in A} \tfrac{1}{n \cdot \binom{n-1}{|A|-1}} \cdot \nu(A) - \sum_{A \subseteq N, i \notin A} \tfrac{1}{n \cdot \binom{n-1}{|A|}} \nu(A) \\
= \;& \sum_{A \subseteq N, i \notin A} \tfrac{1}{n \cdot \binom{n-1}{|A|}} \cdot \left[\nu(A \cup \{i\}) - \nu(A)\right] \\
= \;& \phi_i \qquad \square
\end{aligned}$$

B. Analytical Solution to the Optimization Problem

In order to solve the optimization problem presented in Equation (8), one may use a trick to remove the constraints. One may include both $\emptyset$ and N, as well as $\nu(\emptyset)$ and $\nu(N)$, into the objective and assign them with large weights (e.g., $w_\emptyset = w_N = 10^6$). As a consequence, one ensures that both constraints

$$\nu(\emptyset) \;=\; \sum_{\substack{B\subseteq N\\ |B|\leq k}} \gamma_0^{|B|} I^k(B) \;\text{ and }\; \nu(N) = \sum_{\substack{B\subseteq N\\ |B|\leq k}} \gamma_{|B|}^{|B|} I^k(B)$$

are satisfied when minimizing the objective which implies the constraint $\nu(N) - \nu(\emptyset) = \nu_k(N) - \nu_k(\emptyset)$ of the k-additive optimization problem.

With the aforementioned modifications, the optimization problem can be formulated as follows:

$$\min_{I^k} \quad \sum_{A\in\mathcal{M}} w_A \left(\nu(A) - \sum_{\substack{B\subseteq N\\ |B|\leq k}} \gamma_{|A\cap B|}^{|B|} I^k(B) \right)^2 . \tag{25}$$

Clearly, (25) is a weighted least square problem. Indeed, assume $\mathbf{W}$ as a matrix whose diagonal elements are the weights w_A for all $A \in \mathcal{M}$, $\nu_\mathcal{M}$ as the associated vector of sampled coalitions, and $\mathbf{P}$ as the transformation matrix from the generalized interaction indices to the game, i.e., $\nu_\mathcal{M} = \mathbf{P}I^k$, where $I^k = (I^k(\emptyset), \phi_1^k, \ldots, \phi_n^k, I_{1,2}^k, \ldots, I_{n-1,n}^k, \ldots, I^k(A))$, with $|A| = k$, is the vector of generalized interactions in the lexicographic order for coalitions of players such that $|A| \leq k$. In matrix notation, (25) can be formulated as

$$\min_{I^k} \quad \left(\nu_\mathcal{M} - \mathbf{P}I^k\right)^T \mathbf{W} \left(\nu_\mathcal{M} - \mathbf{P}I^k\right) , \tag{26}$$

whose well-known solution is given by

$$I^k = \left(\mathbf{P}^T\mathbf{W}\mathbf{P}\right)^{-1} \mathbf{P}^T\mathbf{W}\nu_\mathcal{M} . \tag{27}$$

C. Cooperative Games Details

The cooperative games used within our conducted experiments are based on explanation examples for real world data. This section complete their brief description given in Section 5. Across all cooperative games the players represent a fixed set of features given by a particular dataset.

C.1. Global feature importance

Seeking to quantify each feature's individual importance to a model's predictive performance, the value function is based on the model's performance of a hold out test set. This necessitates to split the dataset at hand into training and test set. Features outside of an inspected coalition S are removed by retraining the model on the training set and measuring its performance on the test set. For all games we a applied train-test split of 70% to 30% and a random forest consisting of 20 trees. For classification the value function maps each coalition to the model's resulting accuracy on the test set minus the accuracy of the mode within the data such that the empty coalition has a value of zero. For regression tasks the worth of a coalition is the reduction of the model's mean squared error compared to the empty set which is given by the mean prediction. Again, the empty coalition has a value of zero.

C.2. Local feature attribution

Instead of assessing each feature's contribution to the predictive performance, its influence on a model's prediction for a fixed datapoint can also be investigated. Hence, the value function is based on the model's predicted value.

C.2.1. Adult classification

A sklearn gradient-boosted tree classifies whether a person's annual salary exceeds 50,000 in the *Adult* tabular dataset containing 14 features. The predicted class probability of the true class is taken as the worth of a coalition S. In order to render features outside of S absent, these are imputed by their mean value such that the datapoint is compatible to the model's expected feature number.

C.2.2. Image classification

A *ResNet18* model is used to classify images from *ImageNet*. Since the for error tracking necessary exact computation of Shapley values is infeasible for the given number of pixels, 14 semantic segments are formed after applying *SLIC*. These super-pixels form the player set. Given that the model predicts class c using the full image, the value function assigns to each coalition S the predicted class probability of c resulting from only including those super-pixels in S. The other super-pixels are removed by mean imputation, setting them grey.

C.2.3. IMDB sentiment analysis

A *DistilBERT* transformer fine-tuned on the *IMDB* dataset predicts the sentiment of a natural language sentence between -1 and 1. The sentence is transformed into a sequence of tokens. The input sentences are restricted to sentences that result in 14 tokens being represented by players of the cooperative game. This allows to remove players in the tokenized representation of the transformer. The predicted sentiment is taken as the worth of a coalition.

C.3. Unsupervised feature importance

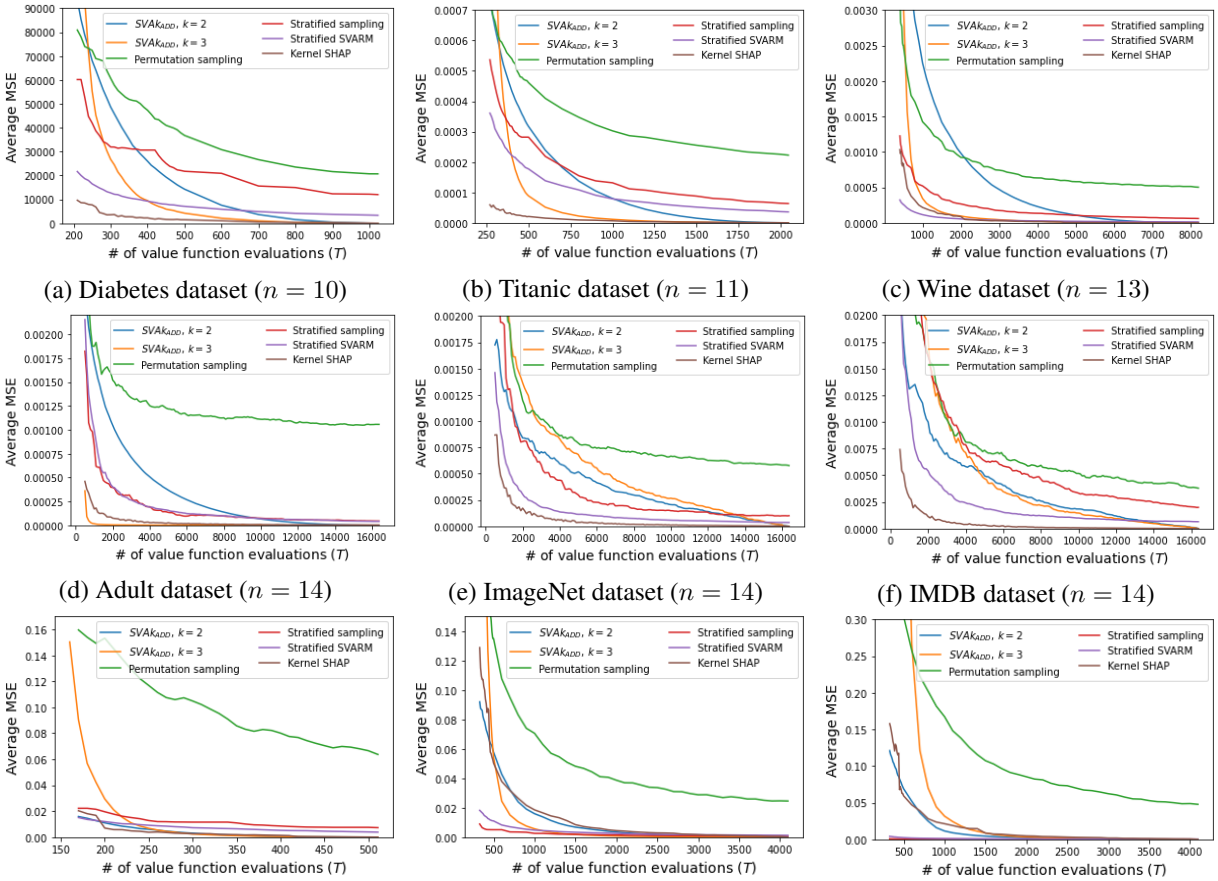
In contrast to the previous settings, there is no available predictive model to investigate unlabeled data. Still, each feature's contribution to the shared information within the data can be quantified and assigned as a score. (Balestra et al., 2022) proposed to view the features $1, \dots, n$ as random variables $X_1, \dots, X_n$ such that the datapoints are realizations of their joint distribution. Next, the worth of a coalition S is given by their total correlation

$$\nu(S) = \sum_{i \in S} H(X_i) - H(S)$$

where $H(X_i)$ denotes the Shannon entropy of X_i and $H(S)$ the contained random variables joint Shannon entropy. The utilized datasets are reduced in the number of features and datapoints to ease computation. The *Breast cancer* dataset contains 9 features and 286 datapoints. The class label indicating the diagnosis is removed. From the *Big five* and *FIFA 21* dataset 12 random features are selected out of the first 50 and the datapoints are reduced to the first 10,000.

D. Further Empirical Results

Aiming at further illustrating the performance of our proposal in comparison with existing approximation methods, in this section, we extend the results presented in Section 5.3 by including another approach, called *ApproShapley* (given here as *Permutation sampling*) (Castro et al., 2009), and our proposal for $k = 2$. We show this comparison in Figure 4. Note that, with the exception of the Wine dataset, the *Permutation sampling* leads to the worst results.



(a) Diabetes dataset ($n = 10$) (b) Titanic dataset ($n = 11$) (c) Wine dataset ($n = 13$)

(d) Adult dataset ($n = 14$) (e) ImageNet dataset ($n = 14$) (f) IMDB dataset ($n = 14$)

(g) Breast Cancer dataset ($n = 9$) (h) Big Five dataset ($n = 12$) (i) FIFA dataset ($n = 12$)

Figure 4: MSE of $SVAk_{\text{ADD}}$ and competing methods averaged over 100 repetitions in dependence of available sample budget T. Datasets stem from various explanation types (i) global (first row), (ii) local (second row), and unsupervised (third row) with differing player numbers n.

C

Appendix to SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification

Organization of the Appendix. Within the Appendix, we provide not only proofs for our theoretical analysis in Appendix E and further empirical results in Appendix G, but also give a table of frequently used symbols throughout the paper inAppendix A, provide Shapley-based interaction measures and other indices falling under the notion of cardinal interaction indices explicitly in Appendix B, provide further and more detailed pseudocode of our algorithmic approach SVARM-IQ in Appendix C, showcase our method at the popular special case of the Shapley Interaction index for pairs Appendix D, and describe the used models, datasets, and explanation tasks within our experimental setup in Appendix F. Appendix H contains the hardware details.

A LIST OF SYMBOLS

Problem setting	
$\mathcal{N}$	Set of players
$\mathcal{N}_k$	Set of all subsets of the player set with cardinality k
n	Number of players
ν	Value function
B	Budget, number of allowed evaluations of ν
k	Considered interaction order
K	Interaction set
$\Delta_K(S)$	Discrete derivative of players K at coalition S
I_K	Cardinal interaction index of players K
$\hat{I}_K$	Estimated cardinal interaction index of players K
$\mathcal{L}_k$	Set of all coalition sizes at which interactions of order k can occur
$\lambda_{k,\ell}$	Weight of each coalition of size ℓ for interaction order k
SVARM-IQ	
$I^W_{K,\ell}$	Average worth of coalitions $S \cup W$ with $S \subseteq \mathcal{N} \setminus K$, $\lvert S\rvert = \ell$ and $W \subseteq K$
$\hat{I}^W_{K,\ell}$	Estimate of $I^W_{K,\ell}$
$c^W_{K,\ell}$	Number of samples observed for stratum $I^W_{K,\ell}$
$\mathcal{S}_{\text{exp}}$	Sizes for which all coalitions are evaluated for explicit stratum computation
$\mathcal{S}_{\text{imp}}$	Sizes for which coalitions are sampled for implicit stratum estimation
$\mathcal{I}_{\text{exp}}$	Set of all explicitly computed strata
$\mathcal{I}_{\text{imp}}$	Set of all implicitly extimated strata
$\mathcal{L}^w_k$	Set of implicit coalitions sizes ℓ depending on W of $I^W_{K,\ell}$
P_k	Probability distribution over sizes $\{2, \ldots, n-2\}$
$\bar{P}_k$	Altered probability distribution over sizes $\{s_{\text{exp}}+1, \ldots, n-s_{\text{exp}}-1\}$
$\bar{B}$	Budget left for the sampling loop after `ComputeBorders`
$\tilde{B}$	Budget left for the sampling loop after `ComputeBorders` and `WarmUp`
$\sigma^2_{K,W,\ell}$	Variance of coalition worths in stratum $I^W_{K,\ell}$
$r_{K,\ell,W}$	Range of coalition worths in stratum $I^W_{K,\ell}$
$\bar{m}^W_{K,\ell}$	Number of samples with which $\hat{I}^W_{K,\ell}$ is updated after the warm-up
$m^W_{K,\ell}$	Total number of samples with which $\hat{I}^W_{K,\ell}$ is updated
$A^W_{K,\ell,m}$	m-th coalition used to update $\hat{I}^W_{K,\ell}$

Table 2: List of symbols used frequently throughout the paper.

B CARDINAL INTERACTION INDICES AND THEIR WEIGHTS

All Shapley-based interaction indices and a few other game-theoretic measures of interaction can be captured under the notion of cardinal interaction indices (CII). We have stated this in Section 2 without presenting the aforementioned indices explicitly. We catch up on this by providing the weights $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$ of each index that is contained within the CII

$$I_K = \sum_{S \subseteq \mathcal{N} \setminus K} \lambda_{k,|S|} \cdot \Delta_K(S)$$

with discrete derivative

$$\Delta_K(S) = \sum_{W \subseteq K} (-1)^{|K|-|W|} \cdot \nu(S \cup W).$$

- Shapley Interaction index (SII) (Grabisch and Roubens, 1999):

$$\lambda^{\text{SII}}_{k,\ell} = \frac{1}{(n-k+1)\binom{n-k}{\ell}}$$

- Shapley-Taylor Interaction index (STI) (Sundararajan et al., 2020):

$$\lambda^{\text{STI}}_{k,\ell} = \frac{k}{n\binom{n-1}{\ell}}$$

- Faithful-Shapley Interaction index (FSI) (Tsai et al., 2023):

$$\lambda^{\text{FSI}}_{k,\ell} = \frac{(2k-1)!}{((k-1)!)^2} \cdot \frac{(n-\ell-1)!(\ell+k-1)!}{(n+k-1)!}$$

- Banzhaf Interaction index (BII) (Grabisch and Roubens, 1999):

$$\lambda^{\text{BII}}_{k,\ell} = \frac{1}{2^{n-k}}$$

For $k = 1$, the SII, STI, and FSI are identical and equal to the Shapley value:

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

which is why these are also called Shapley-based interactions. For a comprehensive overview of the axiomatic background justifying these indices, we refer to (Tsai et al., 2023) and (Fumagalli et al., 2023).

n-SII Values. The n-Shapley Values (n-SII) Φ^n were introduced by Bordt and von Luxburg (2023) as an extension of the Shapley interactions Lundberg et al. (2020) to higher orders. The n-SII constructs an interaction index for interactions up to size n, which is efficient, i.e. the sum of all interactions equals the full model $\nu(\mathcal{N})$. The n-SII are based on SII, I^{SII}, and aggregate SII up to order n. The highest interaction order of n-SII is always equal to SII. For every lower order, the n-SII values are constructed recursively, as

$$\Phi^n_K := \begin{cases} I^{\text{SII}}_K(S) & \text{if } |K| = n \\ \Phi^{n-1}_K + B_{n-|K|} \sum\limits_{\substack{\tilde{K} \subseteq \mathcal{N} \setminus K \\ |K|+|\tilde{K}|=n}} I^{\text{SII}}_{K \cup \tilde{K}} & \text{if } |K| < n, \end{cases}$$

where the initial values are the SV $\Phi^1 \equiv \phi$ and B_n are the Bernoulli numbers. It was shown Bordt and von Luxburg (2023) that n-SII yield an efficient index, i.e.

$$\sum_{\substack{K \subseteq \mathcal{N} \\ |K| \leq n}} \Phi^n_K = \nu(\mathcal{N}).$$

C ADDITIONAL PSEUDOCODE

C.1 Computing Border Sizes

We have only sketched the ComputeBorders procedure and will provide it now in full detail (see Algorithm 2). Its purpose is to determine the coalition sizes $\mathcal{S}_{\text{exp}}$ for which all coalitions are to be evaluated such that the corresponding strata are computed explicitly. We construct this set symmetrically, in the sense that a size s_{exp} is determined such that all $\mathcal{S}_{\text{exp}} = \{0, \ldots, s_{\text{exp}}, n - s_{\text{exp}}, \ldots, n\}$, in other words: the smallest and the largest s_{exp} many set sizes are included. Hence, we assume for simplicity that the initial probability distribution over sizes P_k is symmetric, i.e., $P_k(s) = P_k(n - s)$, although it does not pose a challenge to extend this to any P_k of arbitrary shape.

We start with $s_{\text{exp}} = 1$ and adjust the remaining budget $\bar{B}$ and the altered probability distribution over sizes $\bar{P}_k$. For each size s being included into $\mathcal{S}_{\text{exp}}$, we set its probability mass to zero and upscale the remaining entries, effectively transferring probability mass from the border sizes to the middle. According to this procedure, ComputeBorders constructs $\bar{P}_k$ with

$$\bar{P}_k(s) = \frac{P_k(s)}{\sum_{s' \in \mathcal{S}_{\text{imp}}} P_k(s')} \text{ for all } s \in \mathcal{S}_{\text{imp}} \quad \text{and} \quad \bar{P}_k(s) = 0 \text{ for all } s \in \mathcal{S}_{\text{exp}}.$$

Algorithm 2 ComputeBorders

$s_{\text{exp}} \leftarrow 1$
$\bar{B} \leftarrow B - 2n - 2$
$\bar{P}_k(0), \bar{P}_k(1), \bar{P}_k(n-1), \bar{P}_k(n) \leftarrow 0$
$\bar{P}_k(s) \leftarrow \frac{P_k(s)}{1 - P_k(0) - P_k(1) - P_k(n-1) - P_k(n)}$ for all $s \in \{2, \ldots, n-2\}$
while $s_{\text{exp}} + 1 \leq \frac{n}{2}$ **and** $\binom{n}{s_{\text{exp}}+1} \leq \bar{P}_k(s_{\text{exp}} + 1) \cdot \bar{B}$ **do**
  $s_{\text{exp}} \leftarrow s_{\text{exp}} + 1$
  if $s_{\text{exp}} = \frac{n}{2}$ **then**
    $\bar{B} \leftarrow \bar{B} - \binom{n}{s_{\text{exp}}}$
    $\bar{P}_k \leftarrow \text{Unif}(0, n)$
  else if $s_{\text{exp}} = \frac{n-1}{2}$ **then**
    $\bar{B} \leftarrow \bar{B} - 2\binom{n}{s_{\text{exp}}}$
    $\bar{P}_k \leftarrow \text{Unif}(0, n)$
  else
    $\bar{B} \leftarrow \bar{B} - 2\binom{n}{s_{\text{exp}}}$
    $\bar{P}_k(s_{\text{exp}}) \leftarrow 0$
    $\bar{P}_k(n - s_{\text{exp}}) \leftarrow 0$
    $\bar{P}_k(s) \leftarrow \frac{\bar{P}_k(s)}{1 - 2\bar{P}(s_{\text{exp}})}$ for all $s \in \{s_{\text{exp}} + 1, \ldots, n - s_{\text{exp}} - 1\}$
  end if
end while
$\mathcal{S}_{\text{exp}} \leftarrow \{0, \ldots, s_{\text{exp}}, n - s_{\text{exp}}, \ldots, n\}$
$\mathcal{S}_{\text{imp}} \leftarrow \{s_{\text{exp}} + 1, \ldots, n - s_{\text{exp}} - 1\}$
for $s \in \mathcal{S}_{\text{exp}}$ **do**
  for $A \in \mathcal{N}_s$ **do**
    $v \leftarrow \nu(A)$
    for $K \in \mathcal{N}_k$ **do**
      $W \leftarrow A \cap K$
      $\ell \leftarrow s - |W|$
      $\hat{I}^W_{K,\ell} \leftarrow \hat{I}^W_{K,\ell} + \frac{v}{\binom{n-k}{\ell}}$
    end for
  end for
end for
Output: $\mathcal{S}_{\text{exp}}, \mathcal{S}_{\text{imp}}$

ComputeBorders iterates over sizes in increasing manner, checking whether the reminaing budget $\bar{B}$ is large enough such that the number of coalitions of the next size $s_{\text{exp}}+1$ considered is covered by the expected number of drawn coalitions with that size. As long as this holds true, s_{exp} is incremented and $\bar{B}$ as well as $\bar{P}_k$ are adjusted. Note that thus not only $s_{\text{exp}}+1$ is added to $\mathcal{S}_{\text{exp}}$ but also $n-s_{\text{exp}}-1$. We distinguish between different cases, depending on whether the incremented s_{exp} has reached the middle of the range of coalition sizes. In case of even n this is $\frac{n}{2}$, otherwise $\frac{n-1}{2}$. As soon as s_{exp} reaches that number, $\bar{P}_k(s)$ becomes irrelevant because then all coalitions of all sizes are being evaluated, leaving no strata to be estimated. In this case we simply set $\bar{P}_k$ to the uniform distribution such that it is well-defined.

After the computation of $\mathcal{S}_{\text{exp}}$, we evaluate all coalitions with cardinality $s \in \mathcal{S}_{\text{exp}}$. For each such coalition A we update the estimate $\hat{I}_{K,\ell}^{W}$ with $W = A \cap K$ and $\ell = s - |W|$ according to our update mechanism. Since each stratum contains only coalitions of the same size, this leads to exactly computed strata representing the average of the contained coalitions' worths.

C.2 Warm-up

The WarmUp (see Algorithm 3) procedure guarantees that each stratum estimate $\hat{I}_{K,\ell}^{W}$ with $I_{K,\ell}^{W} \in \mathcal{I}_{\text{imp}}$ is initialized with the worth of one sampled coalition. This is a natural requirement to facilitate our theoretical analysis in Appendix E. We achieve this algorithmically by iterating over all combinations of $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$. Each such combination specifies a stratum that is implicitly to be estimated. WarmUp draws for each stratum a coalition A uniformly at random from the set of all coalitions of size ℓ and not containing any player of K. The estimate $\hat{I}_{K,\ell}^{W}$ is then set to the evaluated worth $\nu(A \cup W)$ and the counter of observed samples is set to one. The spent budget is:

$$
\begin{aligned}
|\mathcal{I}_{\text{imp}}| &= \binom{n}{k} \cdot \sum_{w=0}^{k} \binom{k}{w} |\mathcal{L}_k^{|w|}| \\
&= \binom{n}{k} \cdot \sum_{w=0}^{k} \binom{k}{w} |\{\max\{0, s_{\text{exp}}+1-w\}, \ldots, \min\{n-k, n-s_{\text{exp}}-1-w\}\}| \\
&= \binom{n}{k} \cdot \sum_{w=0}^{k} \binom{k}{w} \left(n - \max\{k, s_{\text{exp}}+1+w\} - \max\{0, s_{\text{exp}}+1-w\} + 1\right).
\end{aligned}
$$

Algorithm 3 WarmUp

for $K \subseteq \mathcal{N}_k$ **do**
  for $W \subseteq K$ **do**
    for $\ell \in \mathcal{L}_k^{|W|}$ **do**
      Draw A from $\{S \subseteq \mathcal{N} \setminus K \mid |S| = \ell\}$ uniformly at random
      $\hat{I}_{K,\ell}^{W} \leftarrow \nu(A \cup W)$
      $c_{K,\ell}^{W} \leftarrow 1$
    end for
  end for
end for

C.3 Updating Strata Mean Estimates

In order to update the mean estimates $\hat{I}_{K,\ell}^{W}$ of the estimated strata incrementally with a single pass, thus not requiring to iterate over all previous samples, we use UpdateMean (see Algorithm 4). Besides the old estimate and the newly observed coalition worth v_b, this requires the number of observations made so far given by $c_{K,\ell}^{W}$.

Algorithm 4 UpdateMean

Input: $\hat{I}_{K,\ell}^{W}$, $c_{K,\ell}^{W}$, v_b
Output: $\frac{\hat{I}_{K,\ell}^{W} \cdot c_{K,\ell}^{W} + v_b}{c_{K,\ell}^{W}+1}$

D THE SPECIAL CASE OF PAIRWISE SHAPLEY-INTERACTIONS

We stated our approximation algorithm SVARM-IQ for all CII and any order k. Since the Shapley Interaction index (SII) for pairs, i.e., $k = 2$, is the most popular among them, we provide a description of SVARM-IQ and the pseudocode (see Algorithm 5) for that specific case, leading to a simpler presentation of our approach.

The SII of a pair of players $\{i, j\} \in \mathcal{N}_2$ is given by

$$I_{i,j}^{\text{SII}} = \sum_{S \subseteq \mathcal{N} \setminus \{i,j\}} \frac{1}{(n-1)\binom{n-2}{|S|}} \left[\nu(S \cup \{i,j\}) - \nu(S \cup \{i\}) - \nu(S \cup \{j\}) + \nu(S)\right].$$

Now, our approach stratifies the discrete derivatives $\Delta_{i,j}(S)$ by size and splits them into multiple strata, yielding the following representation of the SII:

$$I_{i,j}^{\text{SII}} = \frac{1}{n-1} \sum_{\ell=0}^{n-2} I_{i,j,\ell}^{\{i,j\}} - I_{i,j,\ell}^{\{i\}} - I_{i,j,\ell}^{\{i\}} + I_{i,j,\ell}^{\emptyset},$$

with strata terms for all $W \subseteq \{i, j\}$ and $\ell \in \mathcal{L}_2 := \{0, \ldots, n-2\}$:

$$I_{i,j,\ell}^{W} := \frac{1}{\binom{n-2}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \{i,j\} \\ |S| = \ell}} \nu(S \cup W).$$

We keep a stratum estimate $\hat{I}_{i,j,\ell}^{W}$ for each pair i and j, size $\ell \in \mathcal{L}_2$, and subset $W \subseteq \{i, j\}$. Subsequently, the aggregation of the strata estimates, which we obtain during sampling, provides the desired SII estimate:

$$\hat{I}_{i,j}^{\text{SII}} := \frac{1}{n-1} \sum_{\ell=0}^{n-2} \hat{I}_{i,j,\ell}^{\{i,j\}} - \hat{I}_{i,j,\ell}^{\{i\}} - \hat{I}_{i,j,\ell}^{\{i\}} + \hat{I}_{i,j,\ell}^{\emptyset}.$$

For each sampled coalition A of size $|A| = a$, the update mechanism needs to distinguish between only 4 cases. For each pair i and j it updates:

- $\hat{I}_{i,j,a-2}^{\{i,j\}}$ if $i, j \in A$,
- $\hat{I}_{i,j,a-1}^{\{i\}}$ if $i \in A$ but $j \notin A$,
- $\hat{I}_{i,j,a-1}^{\{j\}}$ if $j \in A$ but $i \notin A$, or
- $\hat{I}_{i,j,a}^{\emptyset}$ if $i, j \notin A$.

This case distinction is still captured by computing $W = A \cap K$, $\ell = a - |W|$, and updating $\hat{I}_{i,j,\ell}^{W}$.

Algorithm 5 SVARM-IQ (for the Shapley Interaction index of order $k = 2$)

Input: $(\mathcal{N}, \nu)$, $B \in \mathbb{N}$
$\hat{I}^{\emptyset}_{i,j,\ell}, \hat{I}^{\{i\}}_{i,j,\ell}, \hat{I}^{\{j\}}_{i,j,\ell}, \hat{I}^{\{i,j\}}_{i,j,\ell} \leftarrow 0$ for all $\{i,j\} \in \mathcal{N}_2, \ell \in \mathcal{L}_2$
$c^{\emptyset}_{i,j,\ell}, c^{\{i\}}_{i,j,\ell}, c^{\{j\}}_{i,j,\ell}, c^{\{i,j\}}_{i,j,\ell} \leftarrow 0$ for all $\{i,j\} \in \mathcal{N}_2, \ell \in \mathcal{L}_2$
ComputeBorders
$\bar{B} \leftarrow B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s}$
for $b = 1, \ldots, \bar{B}$ **do**
  Draw size $a_b \in \mathcal{S}_{\text{imp}} \sim \bar{P}_k$
  Draw A_b from $\{S \subseteq \mathcal{N} \mid |S| = a_b\}$ uniformly at random
  $v_b \leftarrow \nu(A_b)$
  for $\{i,j\} \in \mathcal{N}_2$ **do**
    $W \leftarrow A_b \cap \{i,j\}$
    $\ell \leftarrow a_b - |W|$
    $\hat{I}^{W}_{i,j,\ell} \leftarrow \frac{\hat{I}^{W}_{i,j,\ell} \cdot c^{W}_{i,j,\ell} + v_b}{c^{W}_{i,j,\ell} + 1}$
    $c^{W}_{i,j,\ell} \leftarrow c^{W}_{i,j,\ell} + 1$
  end for
end for
$\hat{I}_{i,j} \leftarrow \frac{1}{n-1} \sum_{\ell=0}^{n-2} \hat{I}^{\{i,j\}}_{i,j,\ell} - \hat{I}^{\{i\}}_{i,j,\ell} - \hat{I}^{\{j\}}_{i,j,\ell} + \hat{I}^{\emptyset}_{i,j,\ell}$ for all $\{i,j\} \in \mathcal{N}_2$
Output: $\hat{I}_{i,j}$ for all $\{i,j\} \in \mathcal{N}_2$

E PROOFS

In the following, we give the proofs to our theoretical results in Section 4. We start by defining and revisiting some helpful notation and stating our assumptions.

Notation:

- Let $\mathcal{L}_k := \{0, \dots, n-k\}$.
- Let $\mathcal{L}_k^{|W|} := \{\ell \in \mathcal{L}_k \mid \ell + |W| \in \mathcal{S}_{\text{imp}}\} = [\max\{0, s_{\text{exp}} + 1 - w\}, \min\{n-k, n-s_{\text{exp}}-1\}]$ for any $W \subseteq K \in \mathcal{N}_k$.
- Let $\tilde{B} = B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s} - |\mathcal{I}_{\text{imp}}|$ be the available budget left for the sampling loop after the completion of ComputeBorders and WarmUp.
- For all $K \in \mathcal{N}_k$ with $\ell \in \mathcal{L}_k$, let $A_{K,\ell}$ be a random set with $\mathbb{P}(A_{K,\ell} = S) = \frac{1}{\binom{n-k}{\ell}}$ for all $S \subseteq \mathcal{N} \setminus K$ with $|S| = \ell$.
- For all $K \in \mathcal{N}_k$ with $W \subseteq K$ and $\ell \in \mathcal{L}_k^w$:
  - Let $\sigma^2_{K,\ell,W} := \mathbb{V}[\nu(A_{K,\ell} \cup W)]$ be the strata variance.
  - Let $r_{K,\ell,W} := \max\limits_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W) - \min\limits_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W)$ be the strata range.
  - Let $\bar{m}^W_{K,\ell} := \#\{b \mid |A_b| = \ell + |W|, A_b \cap K = W\}$ be the number of samples with which $\hat{I}^W_{K,\ell}$ is updated during the sampling loop.
  - Let $m^W_{K,\ell} := \bar{m}^W_{K,\ell} + 1$ be the total number of samples with which $\hat{I}^W_{K,\ell}$ is updated.
  - Let $A^W_{K,\ell,m}$ be the m-th coalition used to update $I^W_{K,\ell}$.
- For all $K \in \mathcal{N}_k$ let $R_K := \sum\limits_{W \subseteq K} \sum\limits_{\ell \in \mathcal{L}_k^{|W|}} r_{K,\ell,W}$.
- Let γ_k be $\gamma_2 := 2(n-1)^2$ for $k = 2$ and $\gamma_k := n^{k-1}(n-k+1)^2$ for $k \geq 3$.

Assumptions:

- $\tilde{B} > 0$
- $n \geq 4$
- $B < 2^n$

The lower bound on the leftover budget $\tilde{B}$ is necessary to ensure the completion of ComputeBorders and WarmUp, and that at least one coalition is sampled during the sampling loop. The assumption on n arises from the fact that ComputeBorders automatically evaluates the worth of all coalitions having size $0, 1, n-1$ or n. Hence, all CII values are computed exactly for $n = 3$. Our considered problem statement becomes trivial for $n \leq 2$. Likewise, in order to avoid triviality, we demand the budget to be lower than the total number of coalitions 2^n. Otherwise, all CII values will be computed exactly by ComputeBorders and the approximation problem vanishes. This allows us to state $\mathcal{S}_{\text{imp}} \neq \emptyset$ and $\mathcal{I}_{\text{imp}} \neq \emptyset$.

E.1 Unbiasedness

Lemma E.1. *All strata estimates $\hat{I}^W_{K,\ell}$ are unbiased, i.e., for all $K \in \mathcal{N}_k$, $W \subseteq K$, $\ell \in \mathcal{L}_k$:*

$$\mathbb{E}\left[\hat{I}^W_{K,\ell}\right] = I^W_{K,\ell}.$$

Proof. The statement trivially holds for all strata explicitly computed by ComputeBorders. Thus, we consider the remaining strata which are estimated via sampling. Fix any $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$. Due to the uniform sampling of eligible coalitions once the size is fixed, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\hat{I}_{K,\ell}^{W} \mid m_{K,\ell}^{W}\right] \\
&= \frac{1}{m_{K,\ell}^{W}} \sum_{m=1}^{m_{K,\ell}^{W}} \mathbb{E}\left[\nu(A_{K,\ell,m}^{W}) \mid m_{K,\ell}^{W}\right] \\
&= \frac{1}{m_{K,\ell}^{W}} \sum_{m=1}^{m_{K,\ell}^{W}} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S|=\ell}} \mathbb{P}(A_{K,\ell,m}^{W} = S \cup W \mid |A_{K,\ell,m}^{W}| = \ell + |W|, A_{K,\ell,m}^{W} \cap K = W) \cdot \nu(S \cup W) \\
&= \frac{1}{m_{K,\ell}^{W}} \sum_{m=1}^{m_{K,\ell}^{W}} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S|=\ell}} \frac{1}{\binom{n-k}{\ell}} \cdot \nu(S \cup W) \\
&= \frac{1}{m_{K,\ell}^{W}} \sum_{m=1}^{m_{K,\ell}^{W}} I_{K,\ell}^{W} \\
&= I_{K,\ell}^{W}.
\end{aligned}
$$

Note that the set $A_{K,\ell,m}^{W}$ has cardinality $\ell + |W|$ and fulfills $A_{K,\ell,m}^{W} \cap K = W$ by definition. Otherwise, it would not be used to update $\hat{I}_{K,\ell}^{W}$. Since `WarmUp` gathers one sample for each stratum estimate, it guarantees $m_{K,\ell}^{W} \geq 1$. Thus the above terms are well defined. Finally, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[\hat{I}_{K,\ell}^{W}\right] &= \sum_{m=1}^{\bar{B}+1} \mathbb{E}\left[\hat{I}_{K,\ell}^{W} \mid m_{K,\ell}^{W} = m\right] \cdot \mathbb{P}(m_{K,\ell}^{W} = m) \\
&= \sum_{m=1}^{\bar{B}+1} I_{K,\ell}^{W} \cdot \mathbb{P}(m_{K,\ell}^{W} = m) \\
&= I_{K,\ell}^{W}.
\end{aligned}
$$

□

Theorem 4.1. *The CII estimates returned by SVARM-IQ are unbiased for all $K \in \mathcal{N}_k$, i.e.,*

$$
\mathbb{E}\left[\hat{I}_K\right] = I_K.
$$

Proof. We have already proven the unbiasedness of all strata estimates with Lemma E.1. Thus, we obtain for all $K \in \mathcal{N}_k$:

$$
\begin{aligned}
\mathbb{E}\left[\hat{I}_K\right] &= \mathbb{E}\left[\sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \hat{I}_{K,\ell}^{W}\right] \\
&= \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \mathbb{E}\left[\hat{I}_{K,\ell}^{W}\right] \\
&= \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot I_{K,\ell}^{W} \\
&= I_K.
\end{aligned}
$$

□

E.2 Sample Numbers

Form now on, we distinguish between the special case of order $k = 2$ and all others $k \geq 3$, allowing us to give tighter bounds for the former. Hence, we introduce γ_k for all $k \geq 2$ with

$$\gamma_2 = 2(n-1)^2 \text{ and } \gamma_k = n^{k-1}(n-k+1)^2 \text{ for all } k \geq 3.$$

Lemma E.2. *The number of samples $\bar{m}^W_{K,\ell}$ collected for the strata estimate $\hat{I}^W_{K,\ell}$ of any fixed player set $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$ collected during the sampling loop is binomially distributed with an expected value of at least*

$$\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] \geq \frac{\tilde{B}}{\gamma_k}.$$

Proof. The number of collected samples during the sample loop, i.e. $\bar{m}^W_{K,\ell}$, is binomially distributed because in each iteration the stratum $\hat{I}^W_{K,\ell}$ has the same probability to be updated and the sampled coalitions are independent of each other across the iterations. The number of iterations is $\tilde{B}$ and the condition for an update of $\hat{I}^W_{K,\ell}$ is that the sampled set A_b fulfills $|A_b| = a_b = \ell + |W|$ and $A_b \cap K = W$. This happens with a probability of:

$$\begin{aligned}
&\mathbb{P}(a_b = \ell + |W|, A_b \cap K = W) \\
&= \mathbb{P}(A_b \cap K = W \mid a_b = \ell + |W|) \cdot \mathbb{P}(a_b = \ell + |W|) \\
&= \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot \bar{P}_k(\ell + |W|).
\end{aligned}$$

Hence, we obtain $\bar{m}^W_{K,\ell} \sim Bin\left(\tilde{B}, \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot \bar{P}_k(\ell + |W|)\right)$. This yields

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] &= \tilde{B} \cdot \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot \bar{P}_k(\ell + |W|) \\
&\geq \tilde{B} \cdot \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_k(\ell + |W|).
\end{aligned}$$

Note that $\bar{P}_k(\ell+|W|) \geq P_k(\ell+|W|)$ holds true for all ℓ and $W \subseteq K$ with $\ell+|W| \in \mathcal{S}_{\text{imp}}$ because for these sizes, from which coalitions are sampled, $\bar{P}_k$ can only gain probability mass in comparison to P_k. More precisely, for all $s \in \mathcal{S}_{\text{imp}}$ we have

$$\bar{P}_k(s) = \frac{P_k(s)}{\sum_{s' \in \mathcal{S}_{\text{imp}}} P_k(s')} \geq \frac{P_k(s)}{\sum_{s' \in \mathcal{S}_{\text{exp}}} P_k(s') + \sum_{s' \in \mathcal{S}_{\text{imp}}} P_k(s')} = P_k(s).$$

We continue to prove our statement for the case of $k = 2$ and any fixed $K = \{i, j\}$ by giving a lower bound for the expected value of $\bar{m}^W_{i,j,\ell}$. Inserting $k = 2$, we can further write

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{i,j,\ell}\right] &\geq \tilde{B} \cdot \frac{\binom{n-2}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_2(\ell + |W|) \\
&= \frac{\tilde{B}}{n(n-1)} \cdot \frac{(\ell+|W|)!}{\ell!} \cdot \frac{(n-\ell-|W|)!}{(n-\ell-2)!} \cdot P_2(\ell + |W|).
\end{aligned}$$

Let

$$f(\ell, w) := \frac{(\ell+w)!}{\ell!} \cdot \frac{(n-\ell-w)!}{(n-\ell-2)!} = \begin{cases} (n-\ell)(n-\ell-1) & \text{if } w = 0 \\ (\ell+1)(n-\ell-1) & \text{if } w = 1 \\ (\ell+1)(\ell+2) & \text{if } w = 2 \end{cases}.$$

In the following, we derive the lower bound $f(\ell, |W|) \cdot P_2(\ell + |W|) \geq \frac{n}{2(n-1)}$ for all $|W| \in \{0, 1, 2\}$ and $\ell \in \mathcal{L}_2^{|W|}$ by distinguishing over different cases of n, ℓ, and $|W|$ and exploiting our tailored distribution P_2.

For odd n, $\ell + |W| \leq \frac{n-1}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{\ell(\ell-1)} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)^2}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For odd n, $\ell + |W| \leq \frac{n-1}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(\ell+1)\ell} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)(n+1)}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For odd n, $\ell + |W| \leq \frac{n-1}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(\ell+2)(\ell+1)} \cdot \frac{n-1}{2(n-3)} = \frac{n-1}{2(n-3)} \geq \frac{n}{2(n-1)}$$

For odd n, $\ell + |W| \geq \frac{n+1}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{(n-\ell)(n-\ell-1)} \cdot \frac{n-1}{2(n-3)} = \frac{n-1}{2(n-3)} \geq \frac{n}{2(n-1)}$$

For odd n, $\ell + |W| \geq \frac{n+1}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(n-\ell-1)(n-\ell-2)} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)(n+1)}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For odd n, $\ell + |W| \geq \frac{n+1}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(n-\ell-2)(n-\ell-3)} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)(n+1)}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For even n, $\ell + |W| \leq \frac{n-2}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{\ell(\ell-1)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n^2(n+2)}{2(n-4)(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even n, $\ell + |W| \leq \frac{n-2}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(\ell+1)\ell} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n(n-2)(n+2)}{2(n-4)(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even n, $\ell + |W| \leq \frac{n-2}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(\ell+2)(\ell+1)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} = \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even n, $\ell + |W| \geq \frac{n}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{(n-\ell)(n-\ell-1)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} = \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even n, $\ell + |W| \geq \frac{n}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(n-\ell-1)(n-\ell-2)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n^2}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even n, $\ell + |W| \geq \frac{n}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(n-\ell-2)(n-\ell-3)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

This allows us to conclude:

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{i,j,\ell}\right] &\geq \tilde{B} \cdot \frac{\binom{n-2}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_2(\ell + |W|) \\
&= \frac{\tilde{B}}{n(n-1)} \cdot \frac{(\ell+|W|)!}{\ell!} \cdot \frac{(n-\ell-|W|)!}{(n-\ell-2)!} \cdot P_2(\ell+|W|) \\
&\geq \frac{\tilde{B}}{n(n-1)} \cdot \frac{n}{2(n-1)} \\
&= \frac{\tilde{B}}{\gamma_2}.
\end{aligned}$$

Next, we turn our attention to the case of $k \geq 3$. Inserting the uniform distribution for P_k, we can write for the expected number of samples:

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] &\geq \tilde{B} \cdot \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_k(\ell + |W|) \\
&= \tilde{B} \cdot \frac{(n-k)!}{n!} \cdot \frac{(\ell+|W|)!}{\ell!} \cdot \frac{(n-\ell-|W|)!}{(n-\ell-k)!} \cdot \frac{1}{n-3} \\
&\geq \tilde{B} \cdot \frac{(n-k)!}{n!} \cdot \frac{1}{n-3}.
\end{aligned}$$

In the following we prove that $\frac{(n-k)!}{n!} \cdot \frac{1}{n-3} \geq \frac{1}{n^{k-1}(n-k+1)^2}$. First, we obtain the equivalent inequality

$$n^{k-1}(n-k+1) \geq (n-3) \prod_{i=n-k+2}^{n} i.$$

Note that we have $n \geq k$ at all times. The inequality obviously holds true for all $k \leq 4$. We prove its correctness for $k \geq 5$ by induction over k. We start with the induction base at $k = 5$:

$$\begin{aligned}
& n^{k-1}(n-k+1) \geq (n-3) \prod_{i=n-k+2}^{n} i \\
\Leftrightarrow\ & n^3(n-4) \geq (n-1)(n-2)(n-3)^2 \\
\Leftrightarrow\ & 5n^3 + 39n \geq 29n^2 + 18.
\end{aligned}$$

The resulting equality is obviously fulfilled by all $n \geq 5$. Next, we conduct the induction step by considering the inequality for $k+1$ with $k \geq 5$:

$$\begin{aligned}
& n^k(n-k) \\
=\ & \frac{n(n-k)}{n-k+1} \cdot n^{k-1}(n-k+1) \\
\geq\ & \frac{n(n-k)}{n-k+1} \cdot (n-3) \prod_{i=n-k+2}^{n} i \\
\geq\ & (n-k+1) \cdot (n-3) \prod_{i=n-k+2}^{n} i \\
=\ & (n-3) \prod_{i=n-k+1}^{n} i.
\end{aligned}$$

With the inequality proven, we finally obtain the desired lower bound for the expectation of $\bar{m}_{K,\ell}^W$:

$$\begin{aligned}\mathbb{E}\left[\bar{m}_{K,\ell}^W\right] &\geq \tilde{B} \cdot \frac{(n-k)!}{n!} \cdot \frac{1}{n-3} \\ &\geq \frac{\tilde{B}}{n^{k-1}(n-k+1)^2} \\ &= \frac{\tilde{B}}{\gamma_k}.\end{aligned}$$

□

Lemma E.3. *The expected inverted total sample number of the strata estimate $\hat{I}_{K,\ell}^W$ for any fixed $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$ is bounded by*

$$\mathbb{E}\left[\frac{1}{m_{K,\ell}^W}\right] \leq \frac{\gamma_k}{\tilde{B}}.$$

Proof. In the following, we apply equation (3.4) in (Chao and Strawderman, 1972), stating

$$\mathbb{E}\left[\frac{1}{X+1}\right] = \frac{1-(1-p)^{m+1}}{(m+1)p} \leq \frac{1}{mp} = \frac{1}{\mathbb{E}[X]},$$

for any binomially distributed random variable $X \sim Bin(m,p)$. Due to `WarmUp` we have $m_{K,\ell}^W = \bar{m}_{K,\ell}^W + 1$, since it guarantees exactly one sample for each stratum. Next, Lemma E.2 allows us to substitute X with $\bar{m}_{K,\ell}^W$ and we obtain:

$$\mathbb{E}\left[\frac{1}{m_{K,\ell}^W}\right] = \left[\frac{1}{\bar{m}_{K,\ell}^W + 1}\right] \leq \frac{1}{\mathbb{E}\left[\bar{m}_{K,\ell}^W\right]} \leq \frac{\gamma_k}{\tilde{B}}.$$

□

E.3 Variance and Mean Squared Error

Lemma E.4. *For any $K \in \mathcal{N}_k$, given the sample numbers $m_{K,\ell}^W$ for all $W \subseteq K$ and $\ell \in \mathcal{L}_k^{|W|}$, the variance of the estimate $\hat{I}_K$ is given by*

$$\mathbb{V}\left[\hat{I}_K \mid \left(m_{K,\ell}^W\right)_{\ell \in \mathcal{L}, W \subseteq K}\right] = \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \cdot \frac{\sigma_{K,\ell,W}^2}{m_{K,\ell}^W}.$$

Proof. First, we split the variance of $\hat{I}_K$ with the help of Bienaymé's identity into the variances of the strata estimates and their covariances. Then we make use of the fact that each sample to update a stratum is effectively drawn uniformly:

$$
\begin{aligned}
&\mathbb{V}\left[\hat{I}_K \mid \left(m_{K,\ell}^W\right)_{\ell\in\mathcal{L},W\subseteq K}\right] \\
&= \mathbb{V}\left[\sum_{\ell=0}^{n-k}\binom{n-k}{\ell}\lambda_{k,\ell}\sum_{W\subseteq K}(-1)^{k-|W|}\cdot\hat{I}_{K,\ell}^W \mid \left(m_{K,\ell}^W\right)_{\ell\in\mathcal{L},W\subseteq K}\right] \\
&= \mathbb{V}\left[\sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}\lambda_{k,\ell}\cdot(-1)^{k-|W|}\cdot\hat{I}_{K,\ell}^W \mid \left(m_{K,\ell}^W\right)_{\ell\in\mathcal{L},W\subseteq K}\right] \\
&= \sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\mathbb{V}\left[\hat{I}_{K,\ell}^W \mid m_{K,\ell}^W\right] \\
&\quad + \sum_{\substack{\ell\in\mathcal{L}_k\\W\subseteq K}}\sum_{\substack{\ell'\in\mathcal{L}_k\\W'\subseteq K\\\ell\neq\ell'\vee W\neq W'}}\binom{n-k}{\ell}\binom{n-k}{\ell'}\lambda_{k,\ell}\lambda_{k,\ell'}\cdot(-1)^{2k-|W|-|W'|}\cdot\operatorname{Cov}\left(\hat{I}_{K,\ell}^W,\hat{I}_{K,\ell'}^{W'} \mid m_{K,\ell}^W, m_{K,\ell'}^{W'}\right) \\
&= \sum_{\ell=0}^{n-k}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\sum_{W\subseteq K}\mathbb{V}\left[\hat{I}_{K,\ell}^W \mid m_{K,\ell}^W\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\mathbb{V}\left[\hat{I}_{K,\ell}^W \mid m_{K,\ell}^W\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\mathbb{V}\left[\frac{1}{m_{K,\ell}^W}\sum_{m=1}^{m_{K,\ell}^W}\nu(A_{K,\ell,m}^W) \mid m_{K,\ell}^W\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\cdot\frac{\sigma_{K,\ell,W}^2}{m_{K,\ell}^W}.
\end{aligned}
$$

The strata estimates $\hat{I}_{K,\ell}^W$ and $\hat{I}_{K,\ell'}^{W'}$ are independent for $W \neq W'$ or $\ell = \ell'$ because each sampled coalition A_b can only be used to update one estimate. Consequently, their covariance is zero. Finally, the variances of the estimates for the explicitly calculated strata are zero and thus eliminated. □

Theorem 4.2. *For any $K \in \mathcal{N}_k$ the variance of the estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by*

$$
\mathbb{V}\left[\hat{I}_K\right] \leq \frac{\gamma_k}{\tilde{B}}\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\sigma_{K,\ell,W}^2.
$$

Proof. We combine the variance of each estimate variance conditioned on the sample numbers given by Lemma E.4 with the bound on the expected inverted total sample numbers given by Lemma E.3:

$$
\begin{aligned}
\mathbb{V}\left[\hat{I}_K\right] &= \mathbb{E}_{\left(m_{K,\ell}^W\right)_{\ell\in\mathcal{L},W\subseteq K}}\left[\mathbb{V}\left[\hat{I}_K \mid \left(m_{K,\ell}^W\right)_{\ell\in\mathcal{L},W\subseteq K}\right]\right] \\
&= \mathbb{E}_{\left(m_{K,\ell}^W\right)_{\ell\in\mathcal{L},W\subseteq K}}\left[\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\cdot\frac{\sigma_{K,\ell,W}^2}{m_{K,\ell}^W}\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\sigma_{K,\ell,W}^2\cdot\mathbb{E}\left[\frac{1}{m_{K,\ell}^W}\right] \\
&\leq \frac{\gamma_k}{\tilde{B}}\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda_{k,\ell}^2\sigma_{K,\ell,W}^2.
\end{aligned}
$$

□

Corollary 4.3. *For any $K \in \mathcal{N}_k$ the mean squared error of the estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by*

$$\mathbb{E}\left[\left(\hat{I}_K - I_K\right)^2\right] \leq \frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.$$

Proof. The bias-variance decomposition allows us to decompose the mean squared error into the bias of $\hat{I}_K$ and its variance. Since we have shown the estimate's unbiasedness in Theorem 4.1, we can reduce it to its variance bounded in Theorem 4.2:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{I}_K - I_K\right)^2\right] &= \left(\mathbb{E}\left[\hat{I}_K - I_K\right]\right)^2 + \mathbb{V}\left[\hat{I}_K\right] \\
&= \mathbb{V}\left[\hat{I}_K\right] \\
&\leq \frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.
\end{aligned}$$

□

E.4 Threshold Exceedence Probability

Corollary 4.4. *For any $K \in \mathcal{N}_k$ and fixed $\varepsilon > 0$ the absolute error of the estimate $\hat{I}_K$ returned by SVARM-IQ exceeds ε with a probability of at most*

$$\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \leq \frac{\gamma_k}{\varepsilon^2 \tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.$$

Proof. We apply Chebychev's inequality and make use of the variance bound in Theorem 4.2:

$$\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \leq \frac{\mathbb{V}\left[\hat{I}_k\right]}{\varepsilon^2} \leq \frac{\gamma_k}{\varepsilon^2 \bar{B}_k} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.$$

□

Lemma E.5. *For the stratum estimate $\hat{I}_{K,\ell}^W$ of any $K \in \mathcal{N}_k$ with $W \subseteq K$, $\ell \in \mathcal{L}_k^{|W|}$, and some fixed $\varepsilon > 0$ holds*

$$\mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon \mid m_{K,\ell}^W\right) \leq 2\exp\left(-\frac{2m_{K,\ell}^W \varepsilon^2}{r_{K,\ell,W}^2}\right).$$

Proof. We combine Hoeffding's inequality with the unbiasedness of the strata estimates shown in Lemma E.1 and obtain:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon \mid m_{K,\ell}^W\right) \\
&= \mathbb{P}\left(|\hat{I}_{K,\ell}^W - \mathbb{E}[\hat{I}_{K,\ell}^W]| \geq \varepsilon \mid m_{K,\ell}^W\right) \\
&= \mathbb{P}\left(\left|\sum_{m=1}^{m_{K,\ell}^W} \nu(A_{S,\ell,m}^W) - \mathbb{E}\left[\sum_{m=1}^{m_{K,\ell}^W} \nu(A_{K,\ell,m}^W)\right]\right| \geq m_{K,\ell}^W \varepsilon \mid m_{K,\ell}^W\right) \\
&\leq 2\exp\left(-\frac{2m_{K,\ell}^W \varepsilon^2}{r_{K,\ell,W}^2}\right).
\end{aligned}$$

□

Lemma E.6. *For any $K \in \mathcal{N}_k$ with $W \subseteq K$, $\ell \in \mathcal{L}_k^{|W|}$ and some fixed $\varepsilon > 0$ holds*

$$\mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon\right) \leq \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right) - 1}.$$

Proof. We start by deriving with Hoeffding's inequality and Lemma E.2 a bound on the probability that $\bar{m}_{K,\ell}^W$ falls below $\frac{\tilde{B}}{2\gamma_k}$:

$$\begin{aligned}
&\mathbb{P}\left(\bar{m}_{K,\ell}^W \leq \frac{\tilde{B}}{2\gamma_k}\right) \\
&= \mathbb{P}\left(\mathbb{E}\left[\bar{m}_{K,\ell}^W\right] - \bar{m}_{K,\ell}^W \geq \mathbb{E}\left[\bar{m}_{K,\ell}^W\right] - \frac{\tilde{B}}{2\gamma_k}\right) \\
&\leq \exp\left(-\frac{2\left(\mathbb{E}\left[\bar{m}_{K,\ell}^W\right] - \frac{\tilde{B}}{2\gamma_k}\right)^2}{\tilde{B}}\right) \\
&\leq \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right).
\end{aligned}$$

Further, we show with Lemma E.5 another statement:

$$\begin{aligned}
&\sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\tilde{B}+1} \mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon \mid m_{K,\ell}^W = m\right) \\
&\leq 2 \sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\tilde{B}+1} \exp\left(-\frac{2m\varepsilon^2}{r_{K,\ell,W}^2}\right) \\
&= 2 \sum_{m=0}^{\tilde{B}+1} \exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^m - 2 \sum_{m=0}^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor} \exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^m \\
&= 2\frac{\exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor} - \exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\tilde{B}+1}}{\exp\left(\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right) - 1} \\
&\leq 2\frac{\exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right) - 1}.
\end{aligned}$$

Finally, we combine both intermediate results and obtain:

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon\right) \\
&\leq \sum_{m=1}^{\tilde{B}+1} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \cdot \mathbb{P}\left(m^W_{K,\ell} = m\right) \\
&= \sum_{m=1}^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \cdot \mathbb{P}\left(m^W_{K,\ell} = m\right) \\
&\quad + \sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\bar{B}_k+1} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \cdot \mathbb{P}\left(m^W_{K,\ell} = m\right) \\
&\leq \mathbb{P}\left(m^W_{K,\ell} \leq \left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor\right) + \sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\tilde{B}+1} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{S,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \\
&\leq \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{r^2_{K,\ell,W}}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{r^2_{K,\ell,W}}\right) - 1}.
\end{aligned}
$$

□

Theorem 4.5. *For any $K \in \mathcal{N}_k$ and fixed $\varepsilon > 0$ the absolute error of the estimate $\hat{I}_K$ exceeds ε with probability of at most*

$$
\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \leq \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2 \lambda^2_{k,\ell} R^2_K}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2 \lambda^2_{k,\ell} R^2_K}\right) - 1}.
$$

Proof. We derive the result by applying Lemma E.6 and utilizing the fact that for all explicitly computed strata $I_{K,\ell}^W \in \mathcal{I}_{\text{exp}}$ holds $\hat{I}_{K,\ell}^W = I_{K,\ell}^W$:

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\left|\sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}\lambda_{k,\ell}(-1)^{k-|W|}\left(\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right)\right| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(\sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}\lambda_{k,\ell}\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}\lambda_{k,\ell}\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \varepsilon\right) \\
&\leq \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\mathbb{P}\left(\binom{n-k}{\ell}\lambda_{k,\ell}\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \frac{\varepsilon r_{K,\ell,W}}{R_K}\right) \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\mathbb{P}\left(\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \frac{\varepsilon r_{K,\ell,W}}{\binom{n-k}{\ell}\lambda_{k,\ell}R_K}\right) \\
&\leq \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2\lambda_{k,\ell}^2R_K^2}\right)^{\left\lfloor\frac{\tilde{B}}{2\gamma_k}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2\lambda_{k,\ell}^2R_K^2}\right)-1}.
\end{aligned}
$$

□

F DESCRIPTION OF MODELS, DATASETS AND EXPLANATION TASKS

We briefly sketched the datasets and models on which our cooperative games, used for the experiments, are built. Hence, we provide further details and sources to allow for reproducibility. Note that the LM, CNN, and SOUM are akin to (Fumagalli et al., 2023).

F.1 Language Model (LM)

We used a pretrained sentiment analysis model for movie reviews. To be more specific, it s a variant of `DistilBert`, fine-tuned on the IMDB dataset, and its python version can be found in the *transformers* API (Wolf et al., 2020) at `https://huggingface.co/lvwerra/distilbert-imdb`. The explanation task is to explain the model's sentiment rating between -1 and 1 for randomly selected instances, where positive model outputs indicate positive sentiment. The features, which are words in this case, are removed on the token level, meaning that tokens of missing values are removed from the input sequence of words, shortening the sentence. Thus, a coalition within a given sentence is given by the sequence containing only the words associated with each each player of that coalition. The value function is given by the model's sentiment rating.

F.2 Vision Transformer (ViT)

The ViT is, similar to the LM, a transformer model. Unlike the LM, the ViT operates on image patches instead of words. The python version of the underlying ViT model can be found in the *transformers* API at `https://huggingface.co/google/vit-base-patch32-384`. It originally consists of 144 32x32 pixel image patches, 12 patches for each column and row. In order to calculate the ground truth values exhaustively via brute force, we cluster smaller input patches together into 3x3 images containing 9 patches in total or into 4x4 images containing 16 patches in total. Patches of a cluster are jointly turned on and off depending on whether the cluster is part of the coalition or not. Players, represented by image patches, that are not present in a coalition are removed on the token level and their token is set to the empty token. The worth of a coalition is the model's predicted class probability for the class which has the highest probability for the grand coalition (the original image with no patches removed) and is therefore within $[0, 1]$.

F.3 Convolutional Neural Network (CNN)

The next local explanation scenario is based on a ResNet18[2] model (He et al., 2016b) trained on ImageNet (Deng et al., 2009). The task is to explain the predicted class probability for randomly selected images from ImageNet (Deng et al., 2009). In order to obtain a player set, we use SLIC (Achanta et al., 2012) to merge single pixels to 14 super-pixels. Each super-pixel corresponds to a player in the resulting cooperative game, and a coalition of players entails the associated super-pixels. Absent super-pixel players are removed by setting the contained pixels to grey (mean-imputation). The worth of a coalition is given by the model's predicted class probability, using only the present super-pixels, for the predicted class of the full image with all super-pixels at hand.

F.4 Sum Of Unanimity Models (SOUM)

We further consider synthetic cooperative games, for which the computation of the ground truth values is feasible within polynomial time. For a given player set $\mathcal{N}$ with n many players, we draw $D = 50$ interaction subsets $S_1, \ldots, S_D \subseteq \mathcal{N}$ uniformly at random from the power set of $\mathcal{N}$. Next, we draw for each interaction subset S_d a coefficient $c_d \in [0, 1]$ uniformly at random. The value function is simply constructed by defining

$$\nu(S) = \sum_{d=1}^{D} c_d \cdot [\![S_d \subseteq S]\!]$$

for all coalitions $S \subseteq \mathcal{N}$. We generate 50 instances of such synthetic games and average the approximation results. To our advantage, this construction yields a polynomial closed-form solution of the underlying CII values (Fumagalli et al., 2023), which allows us to use higher player numbers than in real-world explanation scenarios. For details of the CII computation we refer the interested reader to (Fumagalli et al., 2023).

[2] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html

G FURTHER EMPIRICAL RESULTS

We conducted more experiments than shown in the main part but had to omit them due to space constraints. Besides the approximation curves, comparing SVARM-IQ's approximation quality for the SII, STI, and FSI against current baselines measured by the MSE and Prec@10, we present another type of visualization to demonstrate how SVARM-IQ's performance advantage aids in enriching explanations by including interaction effects.

G.1 Further Results on the Approximation Quality

This section contains more detailed versions of the figures depicted in the main section. We compare the approximation quality of SVARM-IQ against baselines for the SII on the LM and ViT in Figure 6, for SII, STI, and FSI for CNN in Figure 7, and for SOUM in Figure 8.

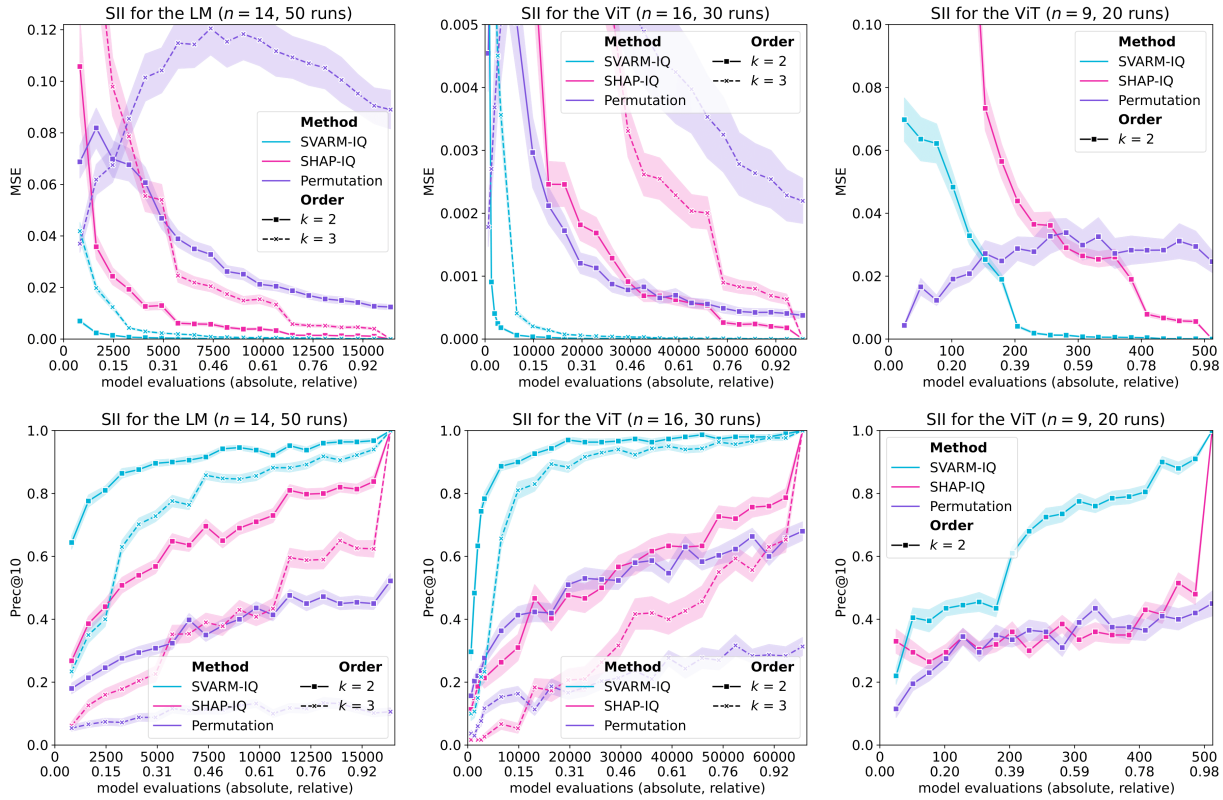


Figure 6: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines averaged over multiple runs for estimating the SII of order $k = 2, 3$ on the LM (first column, $n = 14$, 50 runs) and the ViT (second column, $n = 16$, 30 runs; second column, $n = 9$, 20 runs). The performance is measured by the MSE (first row) and Prec@10 (second row). The shaded bands represent the standard error over the number of performed runs.

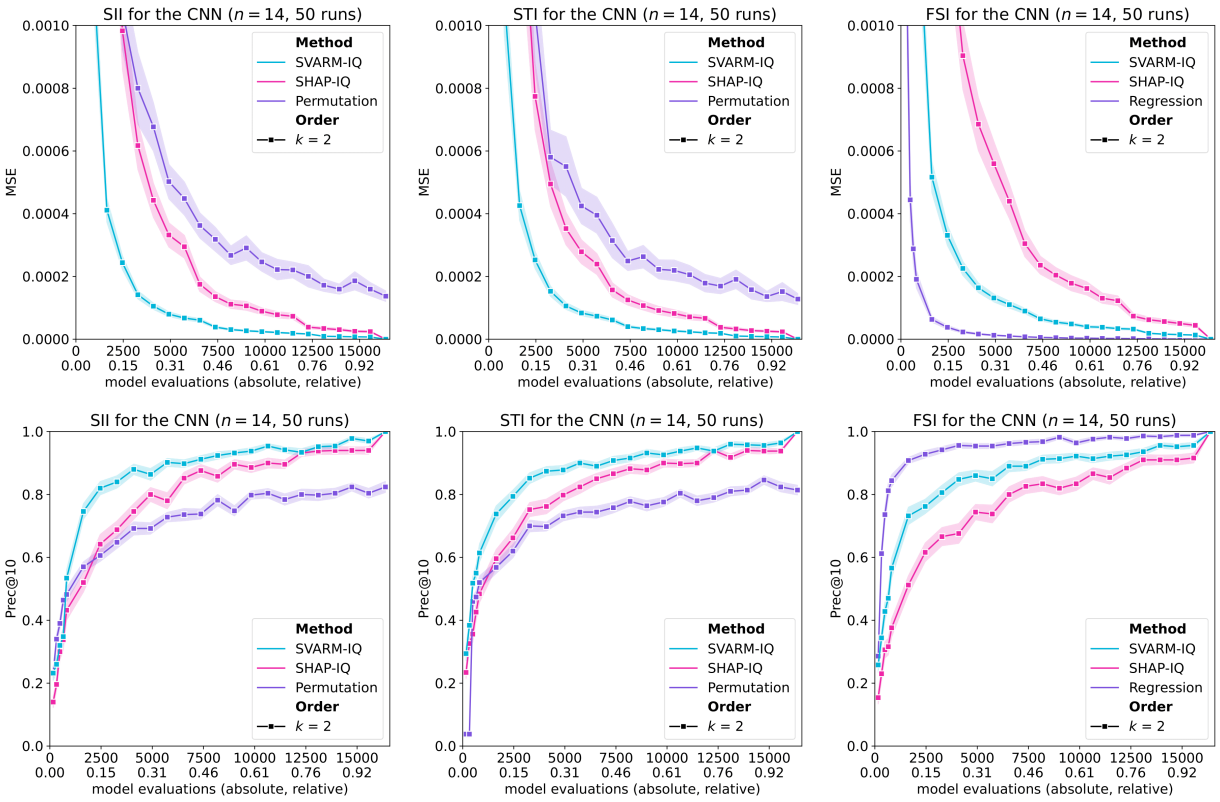


Figure 7: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines averaged over 50 runs on the CNN for estimating the SII (first column), STI (second column), and FSI (third column) of order $k = 2$ for $n = 14$. The performance is measured by the MSE (first row) and Prec@10 (second row). The shaded bands represent the standard error over the number of performed runs.

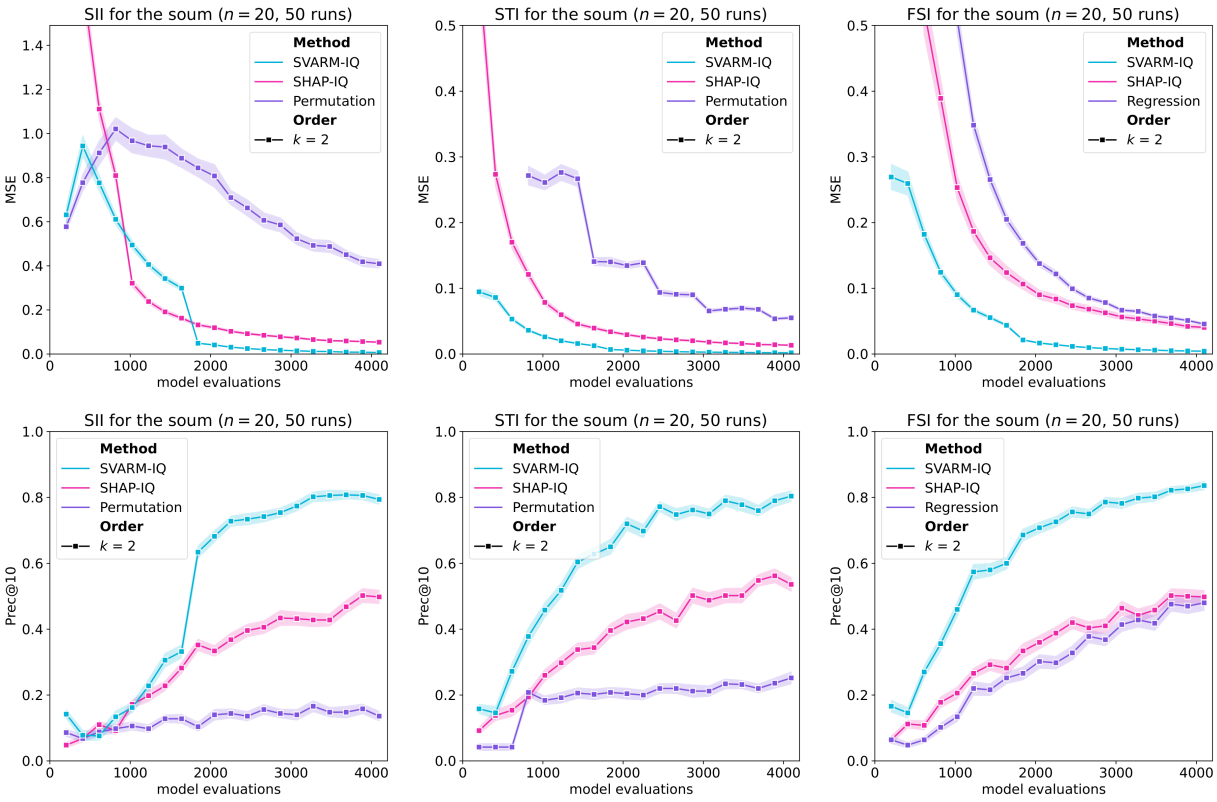


Figure 8: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines averaged over 50 runs on the SOUM for estimating the SII (first column), STI (second column), and FSI (third column) of order $k = 2$ for $n = 20$. The performance is measured by the MSE (first row) and Prec@10 (second row). The shaded bands represent the standard error over the number of performed runs.

G.2 Further Examples of the Vision Transformer Case Study

In the following, we demonstrate how the inclusion of interaction besides attributions scores may enrich interpretability and how significantly SVARM-IQ contributes to more reliable explanations due to faster converging interaction estimates. First, we present in Figure 9 SVARM-IQ's estimates for our ViT scenario, which quantify the importance and interaction of image patches, revealing the insufficiency of sole importance scores and emphasizing the contribution of interaction scores for explaining class predictions for images. Second, we compare in Figure 10 attribution scores and interaction values estimated by SVARM-IQ and permutation sampling with the ground truth. Our results showcase that even with a relatively low number of model evaluations SVARM-IQ mirrors the ground truth almost perfectly, while the inaccurate estimates of its competitor pose the visible risk of misleading explanations, thus harming interpretability.

The obtained estimates for the labrador picture in Figure 9 (upper left) allow for a plausible explanation of the model's reasoning. The most important image patches, those which capture parts of the dogs' heads, share some interesting interaction. The three patches which contain at least one full eye, might be of high importance, but also exhibit strongly negative pairwise interaction. This gives us the insight that the addition of such a patch to an existing one contributes on average little to the predicted class probability in comparison to the increase that such a patch causes on its own, plausibly due to redundant information. In other words, it suffices for the vision transformer to see one patch containing eyes and further patches do not make it much more certain about its predicted class. On the other side, some patches containing different facial parts show highly positive interaction. For example, the teeth and the pair of eyes complement each other since each of them contains valuable information that is missing in the other patch. Considering only the importance scores and their ranking would have not led to this interpretation. Quite the opposite, practitioners would assume most patches to be of equal importance and overlook their insightful interplay.

The comparison of estimates with the ground truth in Figure 10 allows for a twofold conclusion. The estimates obtained by SVARM-IQ show barely any visible difference to the human eye. In fact, SVARM-IQ's approximation replicates the ground truth with only a fraction of the number of model evaluations that are necessary for its exact computation. Hence, it significantly lowers the computational burden for precise explanations. On the contrary, permutation sampling yields estimated importance and interaction scores which are afflicted with evident imprecision. Both, the strength and sign of interaction values are estimated with quite severe deviation for the two considered orders. Hence, the attempt to order the true interactions' strengths or identifying the most influential pairs becomes futile. This lack in approximation quality has the potential to misguide those who seek for explanations on why the model has predicted a certain class.

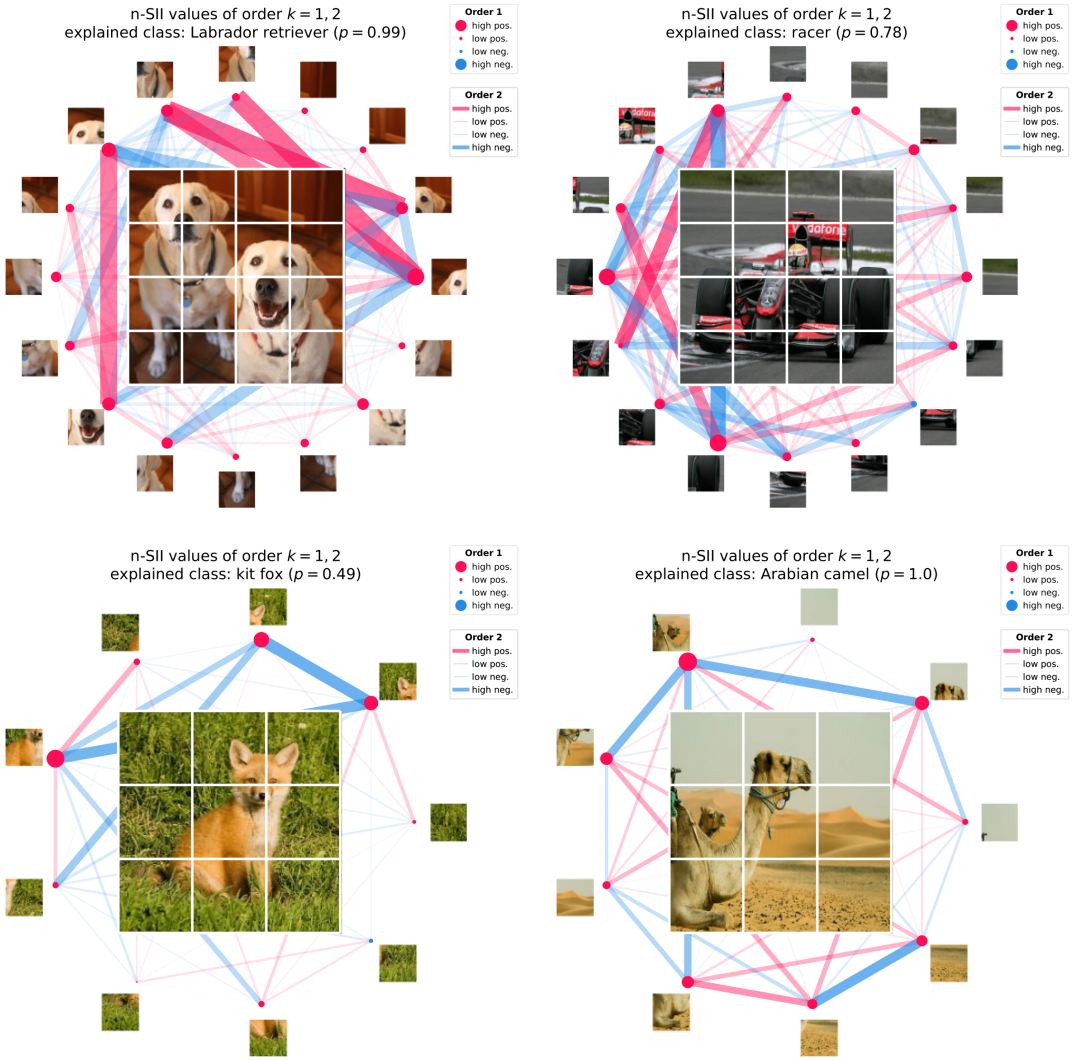


Figure 9: Computed n-SII values of order $k = 1, 2$ by SVARM-IQ for the predicted class probability of a ViT for selected images taken from ImageNet (Deng et al., 2009). The images are sliced into grids of multiple patches, $n = 16$ in the first row and $n = 9$ in the second row. The estimates are obtained after single computation runs given a budget of 10000 evaluations for $n = 16$ patches and 512 (GTV) for $n = 9$ patches.

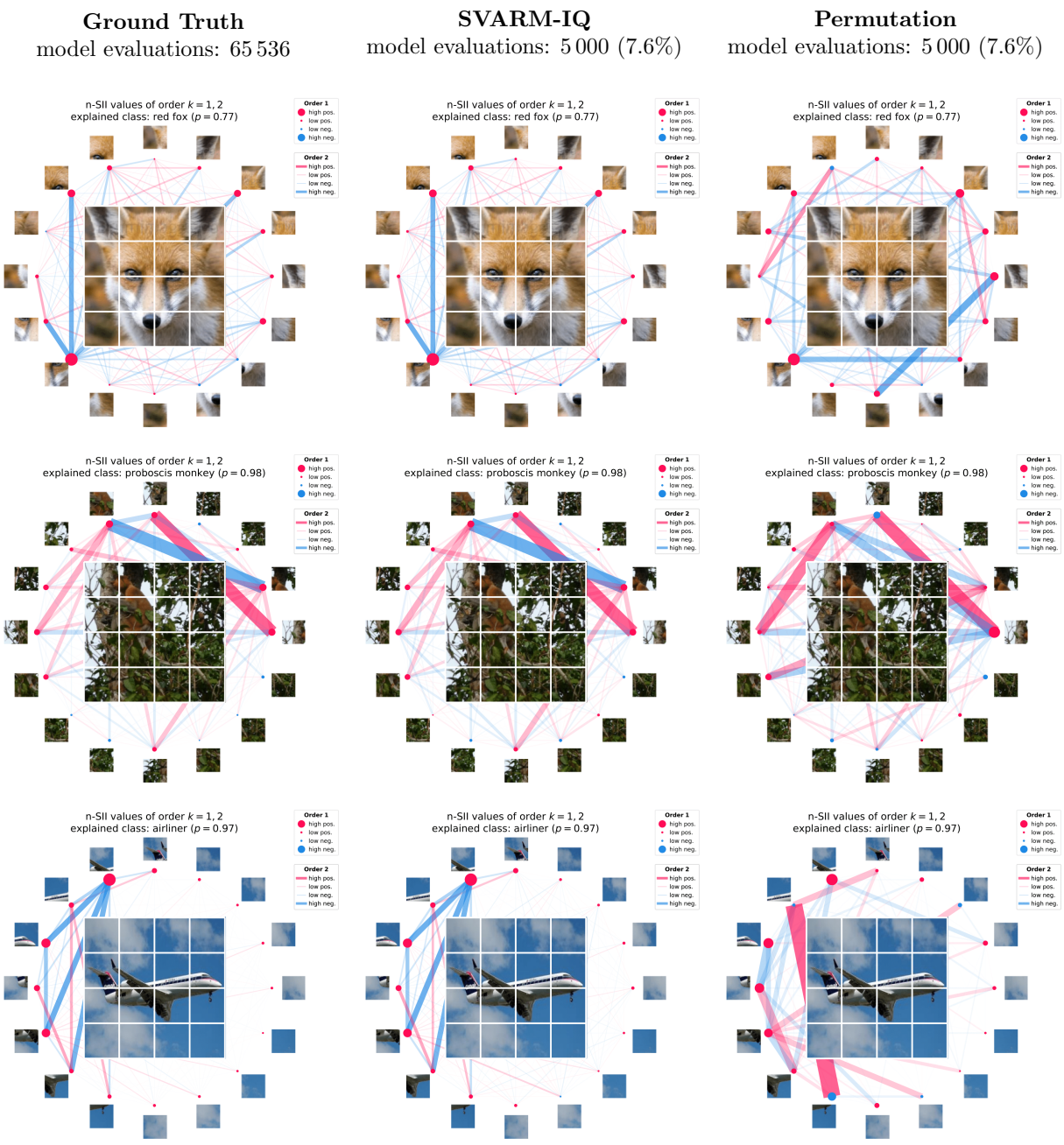


Figure 10: Row-wise comparison of ground-truth n-SII values of order $k = 1, 2$ for the predicted class probability of a ViT (first row) against n-SII values estimated by SVARM-IQ (second column) and permutation sampling (third row) with 5000 model evaluations. The pictures are taken from ImageNet (Deng et al., 2009) and sliced into a grid of 16 patches ($n = 16$).

H HARDWARE DETAILS

This section contains the hardware details required to run and evaluate the empirical results. All experiments where developed and run on a single DELL XPS 15 9510 notebook with Windows 10 Education installed as the operating system. This laptop contains one 11th Gen Intel(R) Core(TM) i7-11800H clocking at 2.30GHz base frequency, 16.0 GB (15.7 GB usable) of RAM, and a NVIDIA GeForce RTX 3050 Ti Laptop GPU.

The model-function calls were pre-computed in around 10 hours on the graphics card. The evaluation of the approximation quality required around 50 hours of work on the CPU. In total, running the experiments took around 50 hours on a single core (no parallelization) and 10 hours on the graphics card.

D

Appendix to Antithetic Sampling for Top-k Shapley Identification

A. Theoretical Analysis

A.1. Proof of Theorem 4.1

For the estimate $\hat{\mathcal{K}} \subseteq \mathcal{N}$ returned by an algorithm for the top-k identification problem we can obviously state

$$\mathbb{P}(\hat{\mathcal{K}} \in \mathbb{K}_\varepsilon) = \sum_{\mathcal{K} \in \mathbb{K}_\varepsilon} \mathbb{P}(\hat{\mathcal{K}} = \mathcal{K})\,.$$

Given the construction of $\hat{\mathcal{K}}$, $\mathcal{A}$ must choose any $i \in \mathcal{N}$ to be in $\hat{\mathcal{K}}$ if $\hat{\phi}_i > \hat{\phi}_j$ holds for at least $n-k$ many players $j \in \mathcal{N}$. Hence, for any $\mathcal{K} \in \mathbb{K}_\varepsilon$ we have:

$$\begin{aligned}
\mathbb{P}(\hat{\mathcal{K}} = \mathcal{K}) \geq\ & \mathbb{P}(\forall i \in \mathcal{K}\, \forall j \in \mathcal{N} \setminus \mathcal{K} : \hat{\phi}_i > \hat{\phi}_j) \\
=\ & 1 - \mathbb{P}(\exists i \in \mathcal{K}\, \exists j \in \mathcal{N} \setminus \mathcal{K} : \hat{\phi}_i \leq \hat{\phi}_j) \\
\geq\ & 1 - \sum_{\substack{i \in \mathcal{K} \\ j \in \mathcal{N} \setminus \mathcal{K}}} \mathbb{P}(\hat{\phi}_i \leq \hat{\phi}_j)
\end{aligned}$$

Given the assumptions on the sampling procedure and the aggregation to estimates $\hat{\phi}_1, \ldots, \hat{\phi}_n$, we can apply the central limit theorem (CLT) to state that for any $i \in \mathcal{K}$ and $j \in \mathcal{N} \setminus \mathcal{K}$ the distribution of $\sqrt{M}\left((\hat{\phi}_i - \hat{\phi}_j) - (\phi_i - \phi_j)\right)$ converges to a normal distribution with mean 0 and variance $\sigma^2_{i,j}$ as $M \to \infty$ since $\mathbb{E}[\hat{\phi}_i - \hat{\phi}_j] = \phi_i - \phi_j$. Although M is finite as it is limited by the budget T, we assume it to be normally distributed, to which it comes close to in practice for large M. Hence, for any $i \in \mathcal{K}$ and $j \in \mathcal{N} \setminus \mathcal{K}$ we derive:

$$\begin{aligned}
\mathbb{P}(\hat{\phi}_i \leq \hat{\phi}_j) =\ & \mathbb{P}(\hat{\phi}_i - \hat{\phi}_j \leq 0) \\
=\ & \mathbb{P}((\hat{\phi}_i - \hat{\phi}_j) - (\phi_i - \phi_j) \leq -(\phi_i - \phi_j)) \\
=\ & \mathbb{P}(\sqrt{M}((\hat{\phi}_i - \hat{\phi}_j) - (\phi_i - \phi_j)) \leq \sqrt{M}(\phi_j - \phi_i)) \\
\overset{CLT}{=}\ & \Phi\left(\sqrt{M}\tfrac{\phi_j - \phi_i}{\sigma_{i,j}}\right)
\end{aligned}$$

where Φ is the standard normal cumulative distribution function. Putting the intermediate results together, we obtain

$$\mathbb{P}(\hat{\mathcal{K}} \in \mathbb{K}_\varepsilon) \geq \sum_{\mathcal{K} \in \mathbb{K}_\epsilon} \left[1 - \sum_{\substack{i \in \mathcal{K} \\ j \in \mathcal{N} \setminus \mathcal{K}}} \Phi\left(\sqrt{M}\tfrac{\phi_j - \phi_i}{\sigma_{i,j}}\right)\right]\,.$$

A.2. Comparable Marginal Contributions Sampling

Proof that Equation (15) induces a well-defined probability distribution:
Obviously it holds $\mathbb{P}(S) \geq 0$ and for the sum of probabilities we have:

$$\sum_{S \subseteq \mathcal{N}} \mathbb{P}(S) = \sum_{S \subseteq \mathcal{N}} \tfrac{1}{(n+1)\binom{n}{|S|}} = \sum_{\ell=0}^{n} \sum_{\substack{S \subseteq \mathcal{N} \\ |S|=l}} \tfrac{1}{(n+1)\binom{n}{\ell}} = \sum_{\ell=0}^{n} \tfrac{\binom{n}{\ell}}{(n+1)\binom{n}{\ell}} = 1\,.$$

Proof of Proposition 5.2:
For any $i \in \mathcal{N}$ we derive:

$$\begin{aligned}
\sum_{S \subseteq \mathcal{N}} \tfrac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta_i'(S) =\ & \sum_{\substack{S \subseteq \mathcal{N} \\ i \in S}} \tfrac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta_i(S \setminus \{i\}) + \sum_{\substack{S \subseteq \mathcal{N} \\ i \notin S}} \tfrac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta_i(S) \\
=\ & \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \tfrac{1}{(n+1)\binom{n}{|S|+1}} \cdot \Delta_i(S) + \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \tfrac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta_i(S) \\
=\ & \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \tfrac{1}{n+1} \left(\tfrac{1}{\binom{n}{|S|+1}} + \tfrac{1}{\binom{n}{|S|}}\right) \cdot \Delta_i(S) \\
=\ & \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \tfrac{1}{n \cdot \binom{n-1}{|S|}} \cdot \Delta_i(S) \\
=\ & \phi_i
\end{aligned}$$

Proof of Proposition 5.3:
Given the unbiasedness of the samples, i.e. $\mathbb{E}[\Delta_i'(S^{(m)})] = \phi_i$ for every $i \in \mathcal{N}$, the covariance is given by:

$$\begin{aligned}\text{Cov}\left(\Delta_i'(S^{(m)}), \Delta_j'(S^{(m)})\right) = &\quad \mathbb{E}\left[\Delta_i'(S^{(m)})\Delta_j'(S^{(m)})\right] - \mathbb{E}\left[\Delta_i'(S^{(m)})\right]\mathbb{E}\left[\Delta_j'(S^{(m)})\right] \\ = &\quad \mathbb{E}\left[\Delta_i'(S^{(m)})\Delta_j'(S^{(m)})\right] - \phi_i\phi_j\end{aligned}$$

For the first term we derive:

$$\begin{aligned}&\mathbb{E}\left[\Delta_i'(S^{(m)})\Delta_j'(S^{(m)})\right] \\ = &\sum_{S\subseteq\mathcal{N}} \frac{1}{(n+1)\binom{n}{|S|}} \cdot \Delta_i'(S)\Delta_j'(S) \\ = &\frac{1}{n+1}\sum_{S\subseteq\mathcal{N}\setminus\{i,j\}} \frac{\Delta_i(S)\Delta_j(S)}{\binom{n}{|S|}} + \frac{\Delta_i(S)\Delta_j(S\cup\{i\})}{\binom{n}{|S|+1}} + \frac{\Delta_i(S\cup\{j\})\Delta_j(S)}{\binom{n}{|S|+1}} + \frac{\Delta_i(S\cup\{j\})\Delta_j(S\cup\{i\})}{\binom{n}{|S|+2}} \\ = &\frac{1}{n+1}\sum_{S\subseteq\mathcal{N}\setminus\{i,j\}} \Delta_i(S)\cdot\left(\frac{\Delta_j(S)}{\binom{n}{|S|}} + \frac{\Delta_j(S\cup\{i\})}{\binom{n}{|S|+1}}\right) + \Delta_i(S\cup\{j\})\cdot\left(\frac{\Delta_j(S)}{\binom{n}{|S|+1}} + \frac{\Delta_j(S\cup\{i\})}{\binom{n}{|S|+2}}\right) \\ = &\frac{1}{n+1}\sum_{S\subseteq\mathcal{N}\setminus\{i\}} \Delta_i(S)\cdot\left(\frac{\Delta_j'(S)}{\binom{n}{|S|}} + \frac{\Delta_j'(S\cup\{i\})}{\binom{n}{|S|+1}}\right)\end{aligned}$$

A.3. Approximating Pairwise Probabilities for Greedy CMCS

Analogously to Appendix A.1, we derive for any pair $i, j \in \mathcal{N}$ and unbiased equifrequent player-wise independent sampler:

$$\begin{aligned}\mathbb{P}(\phi_i < \phi_j) = &\quad \mathbb{P}(\phi_i - \phi_j < 0) \\ = &\quad \mathbb{P}((\hat{\phi}_i - \hat{\phi}_j) - (\phi_i - \phi_j) > \hat{\phi}_i - \hat{\phi}_j) \\ = &\quad \mathbb{P}(\sqrt{M}((\hat{\phi}_i - \hat{\phi}_j) - (\phi_i - \phi_j)) > \sqrt{M}(\hat{\phi}_i - \hat{\phi}_j)) \\ \overset{CLT}{=} &\quad \Phi\left(\sqrt{M}\tfrac{\hat{\phi}_j - \hat{\phi}_i}{\sigma_{i,j}}\right)\end{aligned}$$

Since this statement does not require the knowledge of an eligible coalition $\mathcal{K}$, we can estimate the likelihood of $\phi_i < \phi_j$ during runtime of the approximation algorithm. For this purpose, we use the sample variance to estimate $\sigma_{i,j}$. Note that M is the number of drawn samples that both $\hat{\phi}_i$ and $\hat{\phi}_j$ share. Since the players' marginal contributions are selectively sampled, Greedy CMCS substitutes M by the true number of joint appearances $M_{i,j}$ and $\hat{\phi}_i - \hat{\phi}_j$ by $\hat{\delta}_{i,j}$ which only takes into account marginal contributions of i and j which have been acquired during rounds in which both players have been selected.

B. Pseudocode of Greedy CMCS

In addition to the pseudocode in Algorithm 2, we provide further details regarding the tracking of estimates and probabilistic selection of players.

Algorithm 2 Greedy CMCS

Input: $(\mathcal{N}, \nu), T \in \mathbb{N}, k \in [n], M_{\text{min}}$

$\hat{\phi}_i, M_i \leftarrow 0$ for all $i \in \mathcal{N}$
$M_{i,j}, \Sigma_{i,j}, \Gamma_{i,j} \leftarrow 0$ for all $i, j \in \mathcal{N}$
$t \leftarrow 0$
while $t < T$ **do**
  Draw $\ell \in \{0, \ldots, n\}$ uniformly at random
  Draw $S \subseteq \mathcal{N}$ with $|S| = l$ uniformly at random
  $v_S \leftarrow \nu(S)$
  $t \leftarrow t + 1$
  $P \leftarrow$ SELECTPLAYERS
  for $i \in P$ **do**
    if $t = T$ **then**
      exit
    end if
    $\Delta_i \leftarrow \begin{cases} v_S - \nu(S \setminus \{i\}) & \text{if } i \in S \\ \nu(S \cup \{i\}) - v_S & \text{otherwise} \end{cases}$
    $\hat{\phi}_i \leftarrow \frac{(M_i - 1) \cdot \hat{\phi}_i + \Delta_i}{M_i}$
    $M_i \leftarrow M_i + 1$
    $t \leftarrow t + 1$
  end for
  $M_{i,j} \leftarrow M_{i,j} + 1$ for all $i, j \in P$
  $\Sigma_{i,j} \leftarrow \Sigma_{i,j} + (\Delta_i - \Delta_j)$ for all $i, j \in P$
  $\Gamma_{i,j} \leftarrow \Gamma_{i,j} + (\Delta_i - \Delta_j)^2$ for all $i, j \in P$
end while

Output: $\hat{\mathcal{K}}$ containing k players with highest estimate $\hat{\phi}_i$

- Initialize estimator $\hat{\phi}_i$ and individual counter of sampled marginal contributions M_i for each player.
- Initialize for each player pair: the counter for joint appearances in rounds $M_{i,j}$, the sum of differences of marginal contributions $\Sigma_{i,j}$, and the sum of squared differences of marginal contributions $\Gamma_{i,j}$.
- Given $d_m := \Delta_i(S_m \setminus \{i\}) - \Delta_j(S_m \setminus \{j\})$ the unbiased variance estimator is

$$\hat{\sigma}^2_{i,j} := \frac{1}{M_{i,j}-1} \sum_{m=1}^{M_{i,j}} (d_m - \bar{d})^2 = \frac{1}{M_{i,j}-1} \left(\Gamma_{i,j} - \frac{\Sigma^2_{i,j}}{M_{i,j}} \right) .$$

- In each round, select with SELECTPLAYERS players P for whom to form an extended marginal contribution:
  - First phase: select all players M_{min} times: $P = \mathcal{N}$.
  - Second phase: otherwise, partition the players into top-k players $\hat{\mathcal{K}}$ and the rest $\hat{\mathcal{K}}' = \mathcal{N} \setminus \hat{\mathcal{K}}$ based on the estimates $\hat{\phi}_1, \ldots, \hat{\phi}_n$.
  - Compute $\hat{p}_{i,j} \approx P(\phi_i < \phi_j)$ for all pairs $i \in \hat{\mathcal{K}}, j \in \hat{\mathcal{K}}'$.
  - If all pairs are equally probable, select all players as it is not reasonable to be selective.
  - Otherwise, sample a set of pairs Q based on $\hat{p}_{i,j}$.
  - Select all players as members of P that are in at least one pair in Q.
- Sample a coalition S and cache its value.

- Form for all selected players in P their extended marginal contribution $\Delta_i'(S)$ and update their estimator $\hat{\phi}_i$.
- Update the values $M_{i,j}$, $\Sigma_{i,j}$, and $\Gamma_{i,j}$ for all $i, j \in P$ required for computing the variance estimates $\hat{\sigma}^2_{i,j}$ and $\hat{p}_{i,j}$.
- In practice, we precompute and cache $\nu(\emptyset)$ and $\nu(\mathcal{N})$ in the beginning. We do that for **ALL** tested algorithms for a fair comparison.
- We modify Stratified SVARM to only precompute coalition values for sizes 0 and n, instead of including sizes 1 and $n-1$. Instead of integrating this optimization into all our algorithms, we remove it as it requires a budget of $2n$ which might be infeasible for games with large numbers of players.

Algorithm 3 SELECTPLAYERS

$P \leftarrow \mathcal{N}$
if $M_{i,j} \geq M_{\text{min}}$ for all $i, j \in \mathcal{N}$ **then**
  $\hat{\mathcal{K}} \leftarrow k$ players of $\mathcal{N}$ with highest estimate $\hat{\phi}_i$, solve ties arbitrarily
  $\hat{\mathcal{K}}' \leftarrow \mathcal{N} \setminus \hat{\mathcal{K}}$
  $\hat{\sigma}^2_{i,j} \leftarrow \frac{1}{M_{i,j}-1}\left(\Gamma_{i,j} - \frac{\Sigma^2_{i,j}}{M_{i,j}}\right)$ for all $i \in \hat{\mathcal{K}}, j \in \hat{\mathcal{K}}'$
  $\hat{p}_{i,j} \leftarrow \Phi\left(\sqrt{M_{i,j}}\frac{-\Sigma_{i,j}}{\sqrt{\hat{\sigma}^2_{i,j}}}\right)$ for all $i \in \hat{\mathcal{K}}, j \in \hat{\mathcal{K}}'$
  if $\min_{i,j} \hat{p}_{i,j} \neq \max_{i,j} \hat{p}_{i,j}$ **then**
    $P, Q \leftarrow \emptyset$
    for $(i,j) \in \hat{\mathcal{K}} \times \hat{\mathcal{K}}'$ **do**
      Draw Bernoulli realization $B_{i,j}$ with $\mathbb{P}(B_{i,j} = 1) = \frac{\hat{p}_{i,j} - \min_{i,j} \hat{p}_{i,j}}{\max_{i,j} \hat{p}_{i,j} - \min_{i,j} \hat{p}_{i,j}}$
      if $B_{i,j} = 1$ **then**
        $Q \leftarrow Q \cup \{(i,j)\}$
        $P \leftarrow P \cup \{i,j\}$
      end if
    end for
  end if
end if

Output: $\mathcal{P}$

C. Further Empirical Results

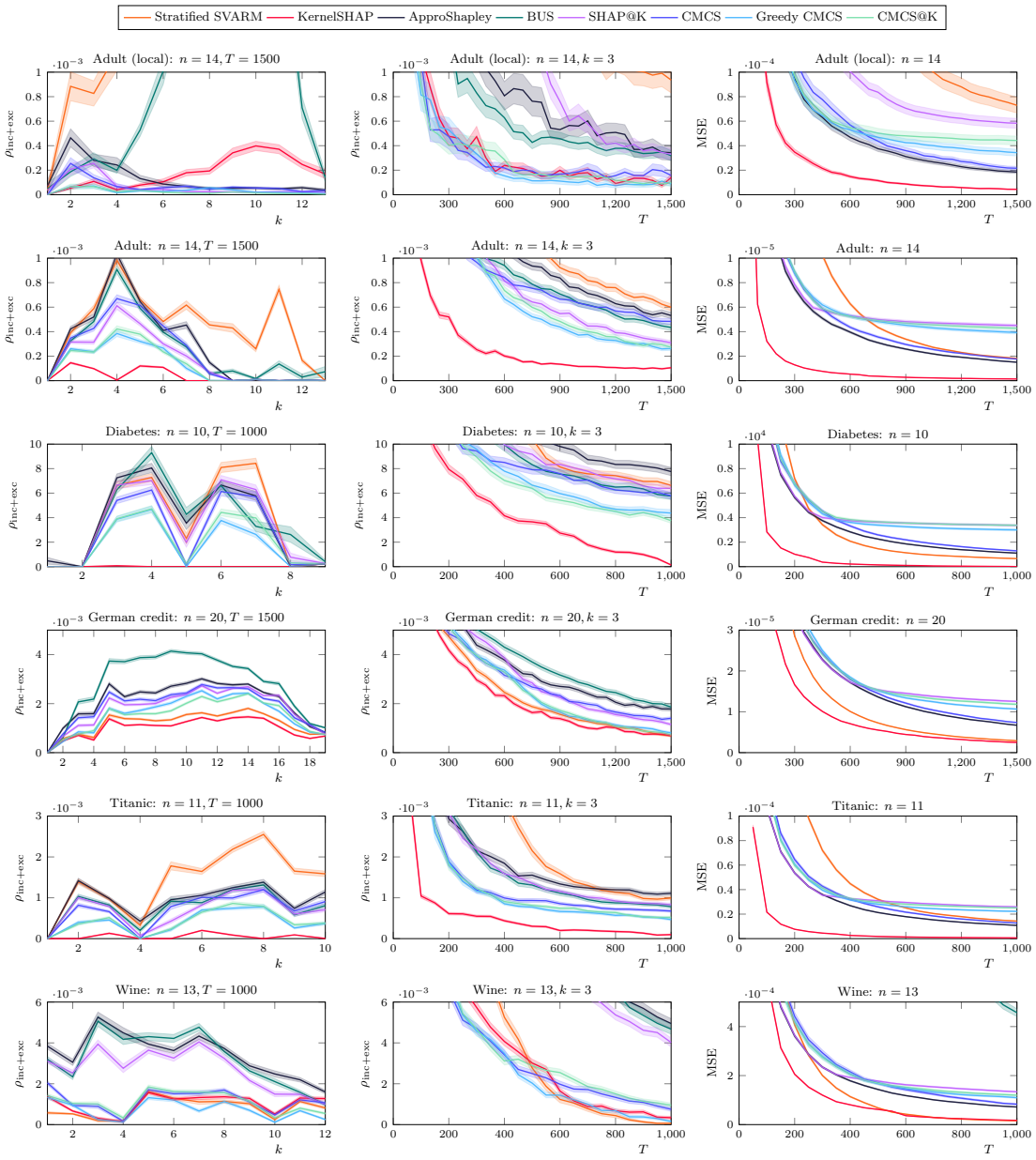


Figure 6. Comparison of achieved error with baselines: inclusion-exclusion error for fixed budget with varying k (left), inclusion-exclusion error for fixed k with increasing budget (middle), and MSE depending on budget (right).

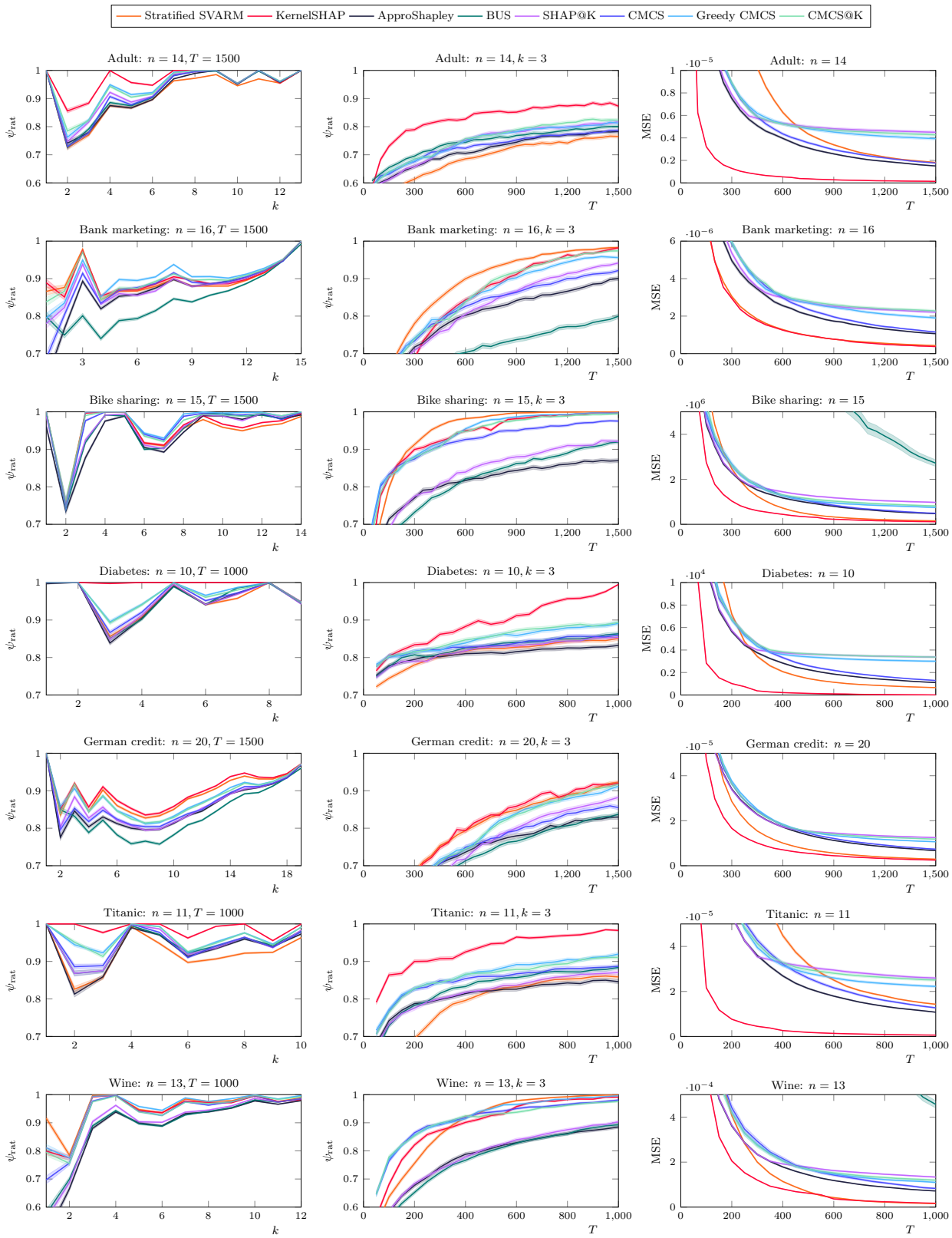


Figure 7. Comparison of achieved ratio precision and MSE with baselines for global explanations: precision for fixed budget with varying k (left), precision for fixed k with increasing budget (middle), and MSE depending on budget (right).

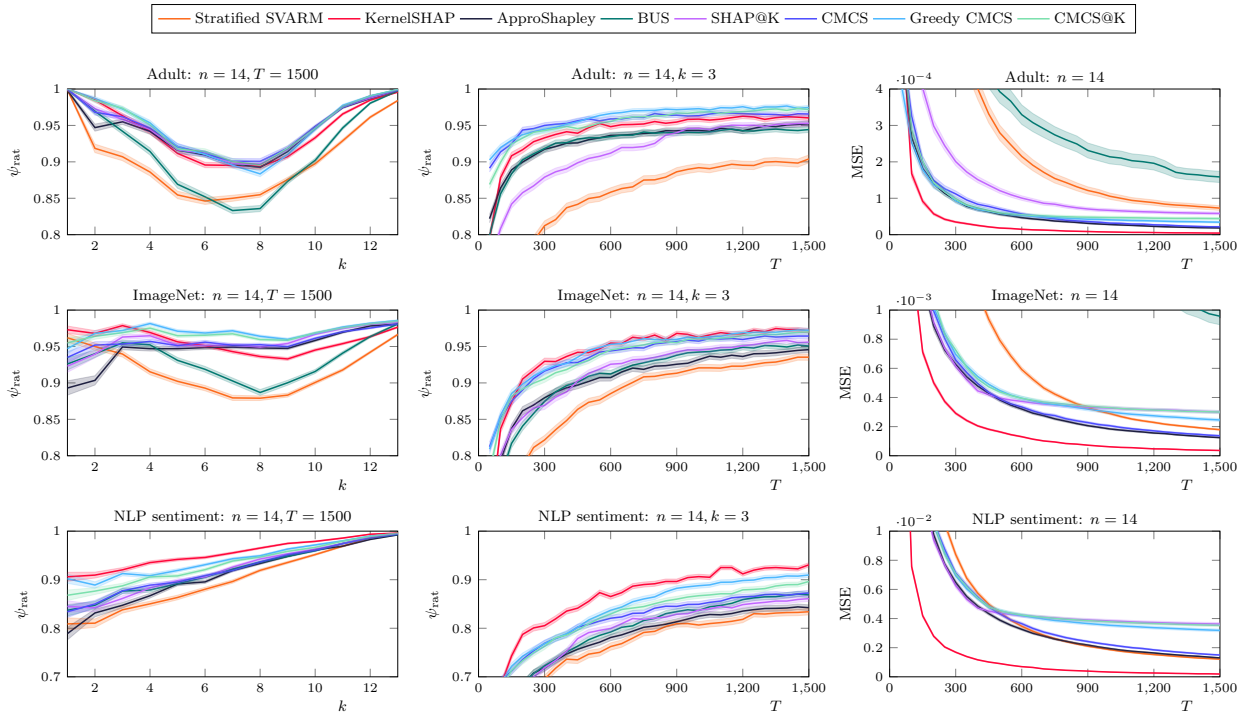


Figure 8. Comparison of achieved ratio precision and MSE with baselines for local explanations: precision for fixed budget with varying k (left), precision for fixed k with increasing budget (middle), and MSE depending on budget (right).

List of Figures

List of Tables

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 08.08.2025

P. Kolpaczki

Patrick Irenäus Kolpaczki