KernelSHAP-IQ: Weighted Least Square Optimization for Shapley Interactions

Fabian Fumagalli [1] **Maximilian Muschalik** [2 3] **Patrick Kolpaczki** [4] **Eyke Hüllermeier** [2 3] **Barbara Hammer** [1]

Abstract

The Shapley value (SV) is a prevalent approach of allocating credit to machine learning (ML) entities to understand black box ML models. Enriching such interpretations with higher-order interactions is inevitable for complex systems, where the Shapley Interaction Index (SII) is a direct axiomatic extension of the SV. While it is well-known that the SV yields an optimal approximation of any game via a weighted least square (WLS) objective, an extension of this result to SII has been a long-standing open problem, which even led to the proposal of an alternative index. In this work, we characterize higher-order SII as a solution to a WLS problem, which constructs an optimal approximation via SII and k-Shapley values (k-SII). We prove this representation for the SV and pairwise SII and give empirically validated conjectures for higher orders. As a result, we propose KernelSHAP-IQ, a direct extension of KernelSHAP for SII, and demonstrate state-of-the-art performance for feature interactions.

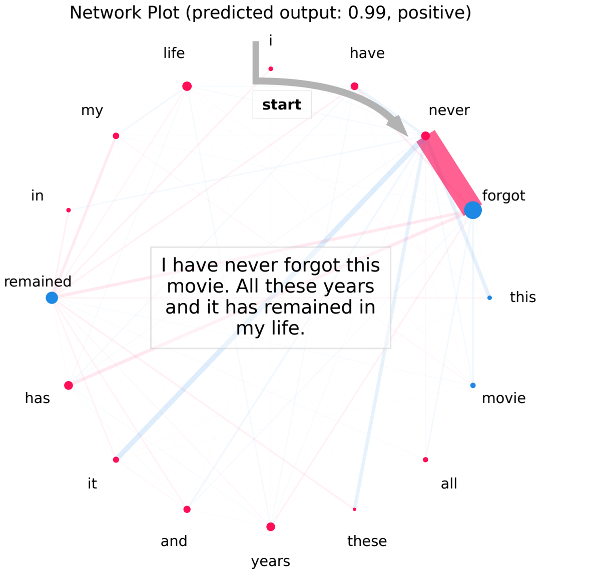

Figure 1. Positive (red) and negative (blue) feature attributions (vertices) and interactions (edges) for a movie review excerpt provided to a sentiment analysis language model. The interaction of "never" and "forget" highly contributes to the positive sentiment.

1. Introduction

The Shapley value (SV) (Shapley, 1953) is a commonly used theoretical framework from cooperative game theory to allocate credit among entities in machine learning (ML) settings. Prevalent ML applications of the SV include *feature attribution* (Lundberg & Lee, 2017) for local interpretability, *feature importance* (Covert & Lee, 2021) for global interpretability, or *data valuation* to quantify the worth of data points (Ghorbani & Zou, 2019). However, in many application domains, the SV accounting solely for individual entities such as features or data points does not suffice (Kumar et al., 2020; 2021; Slack et al., 2020; Sundararajan et al., 2020; Wright et al., 2016). As illustrated with Figures 1 and 2, in complex domains like language modeling (Murdoch et al., 2018), image classification (Tsang et al., 2018), or bioinformatics (Winham et al., 2012; Wright et al., 2016), enriching contributions of individuals with *interactions* among entities is required.

The SV's allocation scheme can be generalized to Shapley interactions and quantify the synergy of groups of entities. The Shapley Interaction Index (SII) (Grabisch & Roubens, 1999) is a natural axiomatic extension of the SV to any-order interactions, which yields an exact additive decomposition of any cooperative game (Grabisch et al., 2000). However, this decomposition includes an exponential number of components and is not suited for interpretability. As a remedy, fixing a maximum interaction order k, k-Shapley values (k-SII) (Bordt & von Luxburg, 2023) aggregate SII up to order k uniquely and yield an *interpretable* and *efficient* k-additive interaction index. This aggregation *fairly* distributes the joint payout, e.g. the model's prediction, to all individuals *and* groups of entities up to order k.

[1]Bielefeld University, CITEC, D-33619 Bielefeld, Germany [2]LMU Munich, D-80539 Munich, Germany [3]MCML, Munich [4]Paderborn University, D-33098, Paderborn, Germany. Correspondence to: Fabian Fumagalli <ffumagalli@techfak.uni-bielefeld.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

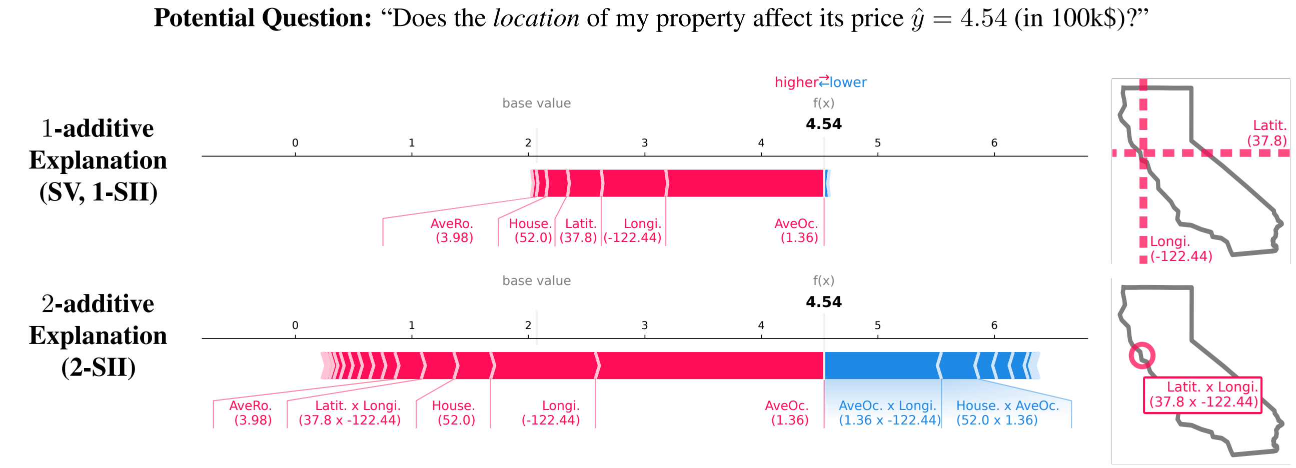

Figure 2. Force plots as first and second order explanations showing positive (red) and negative (blue) feature attributions for a data point of the *California Housing* regression dataset (Kelley Pace & Barry, 1997). The SVs show that a *longitude* of -122.44 and a *latitude* of 37.8, positively influences the predicted property's price. Considering 2-SII feature interactions reveals that the positive influence of *latit.* vanishes and only the combination of *longi.* and *latit.* pointing to the exact location, San Francisco, impacts the property's price.

Both, SII and 1-SII are the SV for individuals. In this case, besides the classical representation, it is well-known that the SV yields an *optimal* 1-additive approximation of the cooperative game, which minimizes the weighted least square (WLS) loss over all subsets constrained on the efficiency property (Charnes et al., 1988). Extensions of such representations for interactions exist for constant weights, which yields the Banzhaf interaction index (Hammer & Holzman, 1992; Grabisch et al., 2000). For Shapley interactions, the Faithful Shapley Interaction Index (FSI) (Tsai et al., 2023) extends on the axiomatic structure of the least square family (Ruiz et al., 1998) by imposing Shapley-like axioms. However, while FSI yields a unique index, it neither yields SII nor k-SII and the dualism between both representations is not as clear as for the SV. In light of the exponential complexity of the SV and SII, which necessitates efficient approximation techniques, a WLS representation of SII is particularly desirable. In fact, KernelSHAP (Lundberg & Lee, 2017) directly utilizes this representation of the SV and yields state-of-the-art (SOTA) approximations in ML applications (Chen et al., 2023; Kolpaczki et al., 2024a). For FSI, where a direct extension of KernelSHAP is possible, it has been shown that it is similarly superior over sampling-based approaches (Fumagalli et al., 2023; Kolpaczki et al., 2024b).

Contribution. In this work, we show that SII is the solution to a WLS problem, which yields the *optimal* k-additive approximation of any cooperative game via k-SII and SII. We then present two variants of KernelSHAP Interaction Quantification (KernelSHAP-IQ), a direct extension of KernelSHAP for higher-order SII. We empirically demonstrate that KernelSHAP-IQ outperforms existing baselines and yields SOTA performance across various datasets and model classes. Therein, our main contributions include:

1. **Iterative approximations via SII**: We introduce k-additive approximations via k-SII and prove an iterative link to SII (Proposition 3.2).
2. **Optimal k-additive approximation via SII**: We prove that pairwise SII are the solution to a WLS problem (Theorem 3.7) with empirically validated conjectures for higher orders (Conjecture 3.9).
3. **KernelSHAP-IQ**[1]: We propose an inconsistent and consistent variant of KernelSHAP for SII that yields SOTA performance for local feature interactions.

Related Work. In cooperative game theory, the studied set functions are known as transferable utility (TU) games, and k-additive TU-games, if k is the maximum order (Grabisch, 2016). Extensions of the SV to interactions were proposed with SII (Grabisch & Roubens, 1999), FSI (Tsai et al., 2023) among others (Marichal & Roubens, 1999; Sundararajan et al., 2020). Recent work was largely motivated by the lack of efficiency of SII, which was recently resolved by the SII-based aggregation k-SII (Bordt & von Luxburg, 2023) that extends on (pairwise) Shapley interactions (Lundberg et al., 2020). The dualism between semivalues, such as the SV or Banzhaf value (Banzhaf III, 1964), and optimal WLS approximations is well known (Ruiz et al., 1998). For interactions, these results extend to Banzhaf interactions (Grabisch & Roubens, 1999; Grabisch et al., 2000). FSI provides a novel index, which directly yields this dualism for Shapley interactions. However, FSI differs from SII and k-SII, and the classical representation, as well as the link between orders remains mostly unknown (Tsai et al., 2023).

[1] KernelSHAP-IQ is implemented in the open-source *shapiq* explanation library `github.com/mmschlk/shapiq`.

Lastly, our approach is linked to the *Shapley residuals* (Kumar et al., 2021).

Approximations of Shapley interactions, were introduced for particular interaction indices by Sundararajan et al. (2020); Tsai et al. (2023) and arbitrary cardinal interaction indices (Fujimoto et al., 2006) by Fumagalli et al. (2023); Kolpaczki et al. (2024b). They directly extend on methods for the SV (Castro et al., 2009; Covert & Lee, 2021; Kolpaczki et al., 2024a). KernelSHAP (Lundberg & Lee, 2017) was introduced for the SV and extended to FSI (Tsai et al., 2023). Including SII in KernelSHAP also improved approximations of the SV (Pelegrina et al., 2023). Above approximations do not make any assumption about the game and were used for model-agnostic interpretability. For tree-based models, it was shown that SII can be computed much more efficiently (Muschalik et al., 2024; Zern et al., 2023), which extends TreeSHAP (Lundberg et al., 2020; Yu et al., 2022) to SII.

Apart from cooperative game theory, interactions were also studied in statistics with functional decomposition (Hooker, 2004; 2007; Lengerich et al., 2020), where some are linked to game theory (Herbinger et al., 2023; Herren & Hahn, 2022; Hiabu et al., 2023). In ML, model-specific variants for deep learning architectures were proposed (Deng et al., 2024; Harris et al., 2022; Janizek et al., 2021; Tsang et al., 2018; 2020a; Zhang et al., 2021), as well as model-agnostic methods (Tsang et al., 2020b) without axiomatic structures.

2. From Shapley Values to Interactions

Notations. We use lower-case letters to denote subset sizes, i.e. $s := |S|$, and write i for $\{i\}$ and ij for $\{i,j\}$. We write constraints in the superscript of a sum, e.g. $\sum_{T\subseteq N}^{k\leq t\leq n-k}$ indicates a sum over subsets $T \subseteq N$ of size $k \leq |T| \leq n-k$. We index matrices using subsets, e.g. $(\mathbf{X})_{TS}$ refers to the row indexed by T and column indexed by S. Lastly, the indicator function $\mathbf{1}_A$ is one if A is fulfilled and $T \sim p$ is a subset sampled according to a probability distribution p.

We consider a set of players $N := \{1,\dots,n\}$ and the payout of a cooperative game $\nu : \mathcal{P}(N) \to \mathbb{R}$, where ν models the payout given a subset of players $T \subseteq N$. In the context of local interpretability, N might be chosen as the set of features and ν the prediction of the model given only a subset of features. For simplicity, we consider games with $\nu(\emptyset) = 0$. In the following, we are interested in quantifying the contribution of a set of players $S \subseteq N$ to the payout ν. For local interpretability, this is known as *feature attribution* for individual players and *feature interaction* for multiple players. For individual players, the *fair* attribution is the well-known Shapley value (SV) (Shapley, 1953)

$$\phi^{\text{SV}}(i) = \sum_{T\subseteq N\setminus i} \tfrac{(n-1-t)!\cdot t!}{n!} \underbrace{[\nu(T\cup i) - \nu(T)]}_{=:\Delta_i(T)}.$$

The SV is the unique attribution measure that fulfills the linearity, symmetry, dummy, and in particular the efficiency axiom, which distributes the payout $\nu(N) - \nu(\emptyset)$ among the players, e.g. the prediction of a ML model. The SV of player i is the weighted average over all marginal contributions $\Delta_i(T) := \nu(T \cup i) - \nu(T)$. For two players ij the marginal contribution to $T \subseteq N \setminus ij$ can be extended using the recursion

$$\Delta_{ij}(T) = \underbrace{\nu(T\cup ij) - \nu(T)}_{\text{contribution of } ij \text{ jointly}} - \underbrace{\Delta_i(T)}_{\text{contribution of } i} - \underbrace{\Delta_j(T)}_{\text{contribution of } j},$$

i.e. the effect of adding both players ij, minus their individual contributions. For feature interactions, a positive value indicates meaningful joint information, e.g. exact position versus latitude and longitude separately, whereas a negative value indicates redundancy. Using above intuition, the marginal contributions can be extended to any subset, which is known as the *discrete derivative* $\Delta_S(T) := \sum_{L\subseteq S}(-1)^{s-l}\nu(T\cup L)$ of a set of players $S \subseteq N$ to a set of participating players $T \subseteq N \setminus S$. Similar to the SV, an interaction index is then defined as a weighted average over discrete derivatives.

Definition 2.1 (Shapley Interaction Index (Grabisch & Roubens, 1999))**.** The Shapley Interaction Index (SII) is

$$\phi^{\text{SII}}(S) := \sum_{T\subseteq N\setminus S} \frac{(n-s-t)!\cdot t!}{(n-s+1)!}\Delta_S(T).$$

We refer to all SIIs of order s as $\phi_s^{\text{SII}} := \{\phi^{\text{SII}}(S) : |S| = s\}$.

The SII includes the SV as $\phi_1^{\text{SII}} \equiv \phi^{\text{SV}}$. It was shown by Grabisch & Roubens (1999) that the SII satisfies the (generalized) linearity, symmetry and dummy axiom. The SII is further the unique interaction index that additionally satisfies the *recursive* axiom, which naturally extends the recursion of discrete derivatives and links higher- to lower-order interactions. For pairwise interaction, it requires

$$\phi(ij) = \underbrace{\phi^{[ij]}([ij])}_{\text{SV of merged player}} - \underbrace{\phi^{-j}(i)}_{\text{SV of } i,\, j \text{ excluded}} - \underbrace{\phi^{-i}(j)}_{\text{SV of } j,\, i \text{ excluded}}.$$

Here, $\phi^{[ij]}, \phi^{-j}, \phi^{-i}$ correspond the SV of games, where $[ij]$ are merged, j is absent, and i is absent, respectively. This recursion is similar to Δ_{ij}, where the right-hand side is the marginal contribution of the merged player $[ij]$ minus the marginal contributions of the individual players, where the other player is absent in T.

2.1. k-Additive Interactions and k-Shapley Values

While the SII is an intuitive extension of the SV to arbitrary interactions, it is not suitable for interpretability as it consists of 2^n non-trivial components. We thus introduce k-additive interaction indices $\Phi_k(S)$, which admit a lower complexity with non-zero interactions up to order $1 \leq |S| \leq k \leq n$, i.e. $\Phi_k(S) := 0$ for $|S| > k$, which we omit for simplicity.

Definition 2.2 (Efficiency)**.** A k-additive interaction index $\Phi_k(S)$ is efficient if and only if $\sum_{S\subseteq N}^{1\leq|S|\leq k} \Phi_k(S) = \nu(N)$.

The SV and the SII subsume all higher order effects, which has been shown by Grabisch (1997) and was recently rediscovered and linked to feature attributions (Bordt & von Luxburg, 2023; Hiabu et al., 2023). When constructing an efficient k-additive interaction index Φ_k based on SII, it is required to account for double counting of effects with order larger than k. For $k = 2$, this was introduced as the (pairwise) Shapley interactions (Lundberg et al., 2020) and extended to higher-order k-Shapley values (Bordt & von Luxburg, 2023).

Definition 2.3 (k-Shapley Values (Bordt & von Luxburg, 2023))**.** The k-Shapley value (k-SII) is defined as $\Phi_k(S) :=$

$$\begin{cases} \phi^{\text{SII}}(S) & \text{if } |S| = k \\ \Phi_{k-1}(S) + B_{k-|S|} \sum_{\tilde{S}\subseteq N\setminus S}^{|S|+\tilde{S}=k} \phi^{\text{SII}}(S \cup \tilde{S}) & \text{if } |S| < k \end{cases}$$

with $1 \leq |S| \leq k \leq n$, $\Phi_0 \equiv 0$ and Bernoulli numbers B_n.

Note that 1-SII is again the SV, i.e. $\Phi_1 \equiv \phi^{\text{SV}}$. Furthermore, k-SII requires SII values up to order k and the highest order is equal to SII. It was shown that Φ_k is efficient and that the Bernoulli numbers are the unique choice of weighting to achieve this property (Bordt & von Luxburg, 2023). For $k = n$ the explicit form is drastically simplified and corresponds to the well-known Möbius transform (Bordt & von Luxburg, 2023). The k-SII therefore defines a flexible low-complexity k-additive interaction index, which is a straight-forward aggregation of SII and can be directly used for an interpretable representation. In Section 3, we will show and leverage that k-SII is directly linked to an *optimal* k-additive approximation of ν.

2.2. Approximations of Shapley Interactions

In applications, the limiting factor of SII and the SV are the amount of evaluations of ν, where both require 2^n evaluations for exact computation. Therefore, multiple approximation techniques have been proposed, as illustrated in Table 1. In the following, we introduce the main approaches for SII, which are based on Monte Carlo sampling of different representations, and extend on techniques for the SV. These methods will be used as baseline methods in our experiments. *Permutation Sampling* (Castro et al., 2009; Tsai et al., 2023) samples random permutations π of N and is based on the representation $\phi^{\text{SII}}(S) = \mathbb{E}_\pi[\mathbf{1}_{\pi\in S}\Delta_S(u_S(\pi))]$, where $u_S(\pi)$ is the set of elements that appear in π before any element of S, and $\mathbf{1}_{\pi\in S}$ is one, if all elements of S appear consecutively in π, and zero otherwise. For permutation sampling only a small fraction of interaction estimates are updated with each evaluation. An alternative method is *SHAP-IQ* (Fumagalli et al., 2023) that samples subsets $T \subseteq N$ directly and utilizes a representation $\phi^{\text{SII}}(S) = \mathbb{E}_T[\nu(T)\omega_s(t, |T \cap S|)]$, where ω_s are SII-specific weights adjusted to the sampling distribution. SHAP-IQ can therefore utilize every evaluation to update all interaction estimates. *SVARM-IQ* (Kolpaczki et al., 2024b) further improves approximation by stratifying over subset sizes t and intersections $T \cap S = L$ with

$$\phi^{\text{SII}}(S) = \sum_{t=0}^{n} \sum_{L\subseteq S} \mathbb{E}[\nu(T)\tilde{\omega}_s(t, L) \mid |T| = t, T \cap S = L],$$

where $\tilde{\omega}_s(t, L) := \omega_s(t, \ell) \cdot p(|T| = t, T \cap S = L)$, i.e. adjusted by the strata probabilities. The core idea is to compute strata estimates by sampling $T \subseteq N$ and treat it as a random sample for one specific stratum of every interaction. This allows again to utilize every evaluation for all interaction estimates. It was shown empirically that SVARM-IQ is the most powerful estimator for SII (Kolpaczki et al., 2024b).

Table 1. Summary of approximation techniques extended to SII

Method	SV	SII (k-SII)
Permutation	(Castro et al., 2009)	(Tsai et al., 2023)
SHAP-IQ	(Covert & Lee, 2021)	(Fumagalli et al., 2023)
SVARM	(Kolpaczki et al., 2024a)	(Kolpaczki et al., 2024b)
KernelSHAP	(Lundberg & Lee, 2017)	**KernelSHAP-IQ**

2.3. Approximating Shapley Values with KernelSHAP

While recent research extended techniques of the SV to the SII, one particularly powerful method has not been considered, so far. KernelSHAP (Lundberg & Lee, 2017) is based on a representation of the SV as a solution to a constrained WLS problem (Charnes et al., 1988; Covert & Lee, 2021)

$$\phi^{\text{SV}} = \underset{\phi\in\mathbb{R}^n}{\arg\min} \sum_{T\subseteq N}^{0<|T|<n} \mu(t) \left[\nu(T) - \sum_{i\in T} \phi(i)\right]^2, \quad \text{such that} \sum_{i\in N} \phi(i) = \nu(N), \tag{1}$$

where $p(T) \propto \mu(t) := \binom{n-2}{t-1}^{-1}$ for $1 \leq t \leq n - 1$.

KernelSHAP considers above objective as an expectation

$$\sum_{T\subseteq N}^{0<t<n} \mu(t)[\nu(T) - \sum_{i\in T} \phi(i)]^2 = \mathbb{E}_T[(\nu(T) - \sum_{i\in T} \phi(i))^2]$$

over sampled subsets $T \sim p(T) \propto \mu(t)$, and approximates the expectation using Monte Carlo sampling. Then, the WLS problem is solved with the approximated objective. The constraint is thereby included with sets $T \in \{\emptyset, N\}$ and a large weight $\mu_\infty \gg 1$, where we prove in Theorem 3.6 that this yields the SV. While the theoretical analysis of KernelSHAP remains difficult, it was shown that the approach works especially well for ML applications (Chen et al., 2023; Kolpaczki et al., 2024a; Lundberg & Lee, 2017). In the following, we rigorously describe this approach and discover such a representation for SII, which we use for efficient approximation.

3. Weighted Least Square Optimization for Shapley Interactions

The SII is an axiomatic extension of the SV and constructs k-SII as an efficient k-additive interaction index used for interpretability. In this section, we give an alternative characterization of SII, as the solution to a WLS problem, which extends on the representation of the SV given in Equation (1). Based on this representation, we introduce KernelSHAP-IQ, an extension of KernelSHAP for Shapley interactions. We rigorously prove its convergence for pairwise interactions and conclude with empirically validated conjectures for higher orders. All proofs are deferred to Appendix A.

The representation of the SV in Equation (1), utilized by KernelSHAP, shows that the approximation $\sum_{i\in T}\phi(i) \approx \nu(T)$ of the game is *optimal* in terms of the WLS objective, if and only if $\phi(i)$ are the SVs. This approximation can be viewed as a 1-additive approximation of the game, which we now directly extend to k-additive approximations.

Definition 3.1 (k-additive Approximation)**.** The k-additive approximation induced by the k-SII Φ_k is given by

$$\hat{\nu}_k(T) := \sum_{S\subseteq T}^{1\leq|S|\leq k} \Phi_k(S) \text{ for } T \subseteq N.$$

Clearly, $\hat{\nu}_1$ is the approximation given in Equation (1). Linking higher orders of the SII to such an optimal approximation of ν has remained unknown. In order to extend KernelSHAP to interactions, recent research introduced the FSI as an alternative index (Tsai et al., 2023). FSI constructs an optimal k-additive approximation of ν by extending on the KernelSHAP representation with interactions up to order k. One disadvantage of FSI is that the solution directly depends on k, which complicates the relationship between representations of different order. Furthermore, the exact representation of lower-order FSI with discrete derivatives remains unknown (Tsai et al., 2023). In contrast, we show that k-SII constructed from SII values yield an appealing iterative approximation of ν, which additively extends on the previous approximation by including a weighted average of SII of order k. We then show that SII via k-SII is linked to an *optimal* k-additive approximation of ν. This yields a suitable framework, where SII is fixed independent of k and k-SII yields an interpretable k-additive interaction index, as summarized in Figure 3.

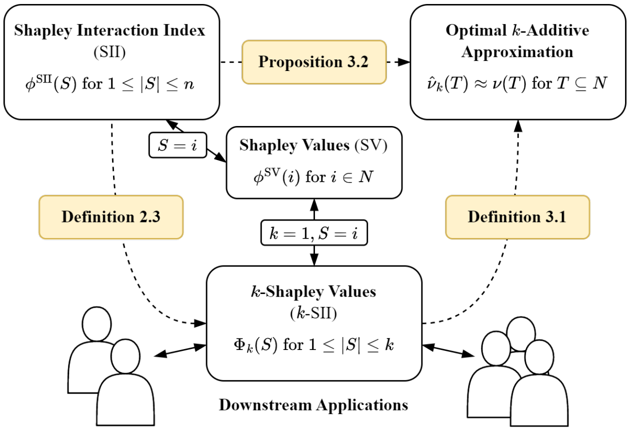
Shapley Interaction Index
(SII)
$\phi^{\text{SII}}(S)$ for $1 \leq |S| \leq n$
Proposition 3.2
Optimal k-Additive
Approximation
$\hat{\nu}_k(T) \approx \nu(T)$ for $T \subseteq N$
$S = i$
Shapley Values (SV)
$\phi^{\text{SV}}(i)$ for $i \in N$
Definition 2.3
Definition 3.1
$k = 1, S = i$
k-Shapley Values
(k-SII)
$\Phi_k(S)$ for $1 \leq |S| \leq k$
Downstream Applications

Figure 3. Links between SII, k-SII and the k-additive approximation $\hat{\nu}_k$. The SII captures the average contribution of S to ν, which constructs the k-additive interaction index k-SII, which is used for interpretation. Both are linked to $\hat{\nu}_k$, where SII yields an *optimal* approximation, which iteratively constructs $\hat{\nu}_k$.

3.1. Optimal Approximations via Shapley Interactions

We first link $\hat{\nu}_k$ and SII, which reveals a surprising recursion.

Proposition 3.2 (Iterative Approximation)**.** *If $k \geq 2$, then*

$$\hat{\nu}_k(T) = \hat{\nu}_{k-1}(T) + \sum_{S\subseteq N}^{|S|=k} \phi^{SII}(S)\lambda(|S|, |S\cap T|)$$

with weights $\lambda(k,\ell) := \sum_{r=1}^{\ell}\binom{\ell}{r}B_{k-r}$ and $\lambda(k,0) := 0$.

Corollary 3.3. *It holds $\lambda(k,k) = 0$ for $k > 1$, $\hat{\nu}_1(T) = \sum_{i\in T}\phi^{SV}(i)$, and $\hat{\nu}_2(T) = \hat{\nu}_1(T) - \frac{1}{2}\sum_{ij\subseteq N}^{|ij\cap T|=1}\phi^{SII}(ij)$.*

Increasing k successively builds on the SV-based approximation $\hat{\nu}_1(T)$ from Equation (1) by including all SII of order k, i.e. $\binom{n}{k}$ components. With increasing k, the approximation $\hat{\nu}_k$ includes more interaction terms, which yields a complexity-accuracy trade-off. In this context, $\hat{\nu}_1$ based on the SVs is the least complex, and $\hat{\nu}_n$ is the most complex and most faithful approximation, where $\hat{\nu}_n(T) = \nu(T)$ follows from a game-theoretic result (Grabisch et al., 2000).

Corollary 3.4. *For any $T \subseteq N$, it holds $\nu(T) = \hat{\nu}_n(T)$.*

In this case, n-SII values ($k = n$ for k-SII), which construct $\hat{\nu}_n$, are the Möbius transform (Bordt & von Luxburg, 2023), which is related to the well-known functional ANOVA decomposition (Hooker, 2004) from statistics, if ν are conditional expectations (Herren & Hahn, 2022).

Optimality of the SII. We now link SII of order k, ϕ_k^{SII}, to an *optimal* k-additive approximation of ν in terms of a

WLS objective. More concretely, we show that for all interaction indices ϕ_k of order k with induced approximation via Proposition 3.2, the SII ϕ_k^{SII} yields the optimal approximation given a specific WLS objective. In the following, it is convenient to use matrix notations, where the rows correspond to subsets $T \subseteq N$ and the columns to interactions $S \subseteq N$ of order $|S| = k$. We first introduce the residual vector $\mathbf{y}_k \in \mathbb{R}^{2^n}$ as $(\mathbf{y}_1)_T := \nu(T)$ and for $k \geq 2$

$$(\mathbf{y}_k)_T := \nu(T) - \hat{\nu}_{k-1}(T) \text{ for } T \subseteq N.$$

Next, we introduce the coefficient matrix $\mathbf{X}_k \in \mathbb{R}^{2^n \times \binom{n}{k}}$ as

$$(\mathbf{X}_k)_{TS} := \lambda(|S|, |T \cap S|) \text{ for } T, S \subseteq N \text{ with } |S| = k.$$

With SII of order k, $\phi_k^{\text{SII}} \in \mathbb{R}^{\binom{n}{k}}$, and Proposition 3.2 the approximation error of ν and $\hat{\nu}_k$ is given by $\mathbf{y}_k - \mathbf{X}_k \phi_k^{\text{SII}}$. Introducing a diagonal weight matrix $\mathbf{W}_k \in \mathbb{R}^{2^n \times 2^n}$, we aim to characterize ϕ_k^{SII} as the solution to a WLS problem

$$\phi^* = \underset{\phi_k \in \mathbb{R}^{\binom{n}{k}}}{\arg\min} \left\| \sqrt{\mathbf{W}_k} \left(\mathbf{y}_k - \mathbf{X}_k \phi_k\right) \right\|_2^2, \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean norm. Note that this WLS problem for $k = 1$ is similar to Equation (1) but includes all subsets $T \subseteq N$. The solution is explicitly given as

$$\phi^* = (\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{W}_k \cdot \mathbf{y}_k. \tag{3}$$

We now define the weights of $\mathbf{W}_k$ with $\mu_\infty \gg 1$ as

$$(\mathbf{W}_k)_{TT} := \mu_k(t) := \begin{cases} \binom{n-2\cdot k}{t-k}^{-1} & \text{if } k \leq t \leq n-k \\ \mu_\infty & \text{else.} \end{cases}$$

Remark 3.5. The weights μ_k appear naturally from $\phi^{\text{SII}}(S) = \sum_{T \subseteq N} \nu(T) \omega_s(t, |T \cap S|)$ (Fumagalli et al., 2023) as the common factors $\mu_k(t) \propto \omega_s(t, 0), \ldots \omega_s(t, k)$ for $k \leq t \leq n - k$. Since $(\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1} \in \mathbb{R}^{\binom{n}{k} \times \binom{n}{k}}$ does not depend on T, it holds that μ_k controls the final weight in $\mathbf{X}_k^T \mathbf{W}_k$ for all interactions S. For a more detailed discussion, see Appendix B.9.

For $k = 1$, μ_1 are the KernelSHAP weights and the SV is the solution of the WLS problem for $\mu_\infty \to \infty$, which justifies KernelSHAP.

Theorem 3.6 (KernelSHAP)**.** *Let $n \geq 2$ and $(\mathbf{W}_1)_{TT} := \mu_1(t)$. Then the SV is represented as*

$$\phi^{SV} = \lim_{\mu_\infty \to \infty} \underset{\phi_1 \in \mathbb{R}^n}{\arg\min} \left\| \sqrt{\mathbf{W}_1} \left(\mathbf{y}_1 - \mathbf{X}_1 \phi_1\right) \right\|_2^2.$$

A large weight μ_∞ requires the solution to achieve lower approximation error for this subset, which is a soft constraint. The subsets with weights μ_∞ are the empty set and N. For the empty set, the approximation error is constant, and is thus not influenced by the solution of the WLS problem. For N, the approximation error is $(\mathbf{y}_1)_N - (\mathbf{X}_1 \phi)_N = \nu(N) - \sum_{i \in N} \phi(i)$, which is zero if efficiency holds. Intuitively, with $\mu_\infty \to \infty$, the soft constraint becomes a hard constraint, which requires zero approximation error, i.e. efficiency with $\nu(N) = \sum_{i \in N} \phi(i)$. This was argued in KernelSHAP (Lundberg & Lee, 2017) without formal proof. Theorem 3.6 is a formal proof that this intuition holds, and indeed yields the SV satisfying the efficiency constraint from Equation (1).

For pairwise SII, we now present a novel representation akin to Theorem 3.6.

Theorem 3.7 (KernelSHAP-IQ, $k = 2$)**.** *Let $n \geq 4$ and $(\mathbf{W}_2)_{TT} := \mu_2(t)$. Then the pairwise SII is represented as*

$$\phi_2^{SII} = \lim_{\mu_\infty \to \infty} \underset{\phi_2 \in \mathbb{R}^{\binom{n}{2}}}{\arg\min} \left\| \sqrt{\mathbf{W}_2} \left(\mathbf{y}_2 - \mathbf{X}_2 \phi_2\right) \right\|_2^2.$$

While $\mu_\infty \to \infty$ corresponds to a constraint for $k = 1$ that ensures efficiency, its behavior for $k = 2$ is less clear. In fact, $\mu_2(1) = \mu_2(n-1) = \mu_\infty$, but for two such subsets $T \in \{\ell, N \setminus \ell\}$ with $\ell \in N$ the approximation is for both the same value as $(X_2\phi_2)_T = -\frac{1}{2} \sum_{ij \subseteq N}^{|T \cap ij| = 1} \phi(ij) = -\frac{1}{2} \sum_{j \in N}^{j \neq \ell} \phi(j\ell)$, which implies that the corresponding hard constraint of zero approximation error cannot be satisfied. Consequently, contrary to $k = 1$, the soft constraints with μ_∞ cannot be re-written to a hard-constrained WLS problem, since it would not have a solution in general.

3.2. Optimal Higher-Order Approximations

For $k > 2$ we were unable to find closed-form solutions and we suspect that our proof may not be suited for finding these in general. However, we empirically validated that μ_k can also be used for higher-order approximations, which we summarize in the following conjectures. The first one concerns the structure of the inverse in Equation (3).

Conjecture 3.8. Let $n \geq 2k$ and define the *precision matrix*

$$\mathbf{A}_k := \lim_{\mu_\infty \to \infty} (\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1},$$

Then, for $S_1, S_2 \subseteq N$ with $|S_1| = |S_2| = k$, it holds

$$(\mathbf{A}_k)_{S_1 S_2} = \frac{(-1)^{k - |S_1 \cap S_2|}}{n - k + 1} \binom{n-k}{k - |S_1 \cap S_2|}^{-1}.$$

Referring to $\mathbf{A}_k$ as the precision matrix is due to the interpretation of $\mathbf{A}_1^{-1}$ as the *covariance matrix* for the SV by Covert & Lee (2021). The second conjecture poses a higher-order representation of SII. In fact, if Conjecture 3.8 holds, then the subsets $T \subseteq N$ with finite weights, i.e. $k \leq |T| \leq n - k$ yield the correct weights for SII.

Conjecture 3.9. Let $n \geq 2k$ and $\mathcal{T}_k := \{T \subseteq N : k \leq |T| \leq n-k\}$, which splits for $T \subseteq N$ the vector $\mathbf{y}_k$ into

$$(\mathbf{y}_k^+)_T := \mathbf{1}_{T \in \mathcal{T}_k} \cdot \mathbf{y}_k(T), \quad (\mathbf{y}_k^-)_T := \mathbf{1}_{T \notin \mathcal{T}_k} \cdot \mathbf{y}_k(T).$$

Let further for $T \subseteq N$ and interaction $S \subseteq N$ with $|S| = k$

$$(\mathbf{Q}_k)_{ST} := \begin{cases} (-1)^{s-|S\cap T|} m_s(t - |S \cap T|), & \text{if } T \notin \mathcal{T}_k, \\ 0, & \text{if } T \in \mathcal{T}_k, \end{cases}$$

where $m_s(t) := \frac{(n-s-t)!\cdot t!}{(n-s+1)!}$ from Definition 2.1. Then,

$$\phi_k^{\text{SII}} = \mathbf{Q}_k \mathbf{y}_k^- + \lim_{\mu_\infty \to \infty} \operatorname*{arg\,min}_{\phi_k \in \mathbb{R}^{\binom{n}{k}}} \left\| \sqrt{\mathbf{W}_k} \left(\mathbf{y}_k^+ - \mathbf{X}_k \phi_k \right) \right\|_2^2 .$$

Conjecture 3.9 states that higher-order SII are again represented as the solution to a WLS problem. However, in contrast to $k \leq 2$, the representation does only hold partially for $T \in \mathcal{T}_k$. For $T \notin \mathcal{T}_k$, $\mathbf{Q}_k$ contains all SII weights (Fumagalli et al., 2023), which are required for the correct weighting. This is necessary as, in contrast to $k \leq 2$, the weights do not converge to the SII weights for $\mu_\infty \to \infty$ and $T \notin \mathcal{T}_k$, which requires to adjust the representation using $\mathbf{y}_k^+$ and $\mathbf{y}_k^-$. Presumably, different μ_∞ could resolve this issue, which we leave to future research.

3.3. KernelSHAP-IQ for Shapley Interactions

We introduce two variants of KernelSHAP-IQ for SII based on distinct optimization problems. Inconsistent KernelSHAP-IQ solves a single WLS problem and does not converge to SII. In contrast, (consistent) KernelSHAP-IQ solves iteratively a WLS problem for every order and converges to SII. Similar to KernelSHAP, we solve an *approximated objective* by sampling subsets from a distribution $p^*(T) \propto q(t)$ with *sampling weights* $q(t) \geq 0$ for $t = 0, \dots, n$. We re-write Equation (2) as

$$\phi^* = \operatorname*{arg\,min}_{\phi_k \in \mathbb{R}^{\binom{n}{k}}} \mathbb{E}_{T \sim p^*} \left[\left\| \sqrt{\mathbf{W}_k^*} \left(\mathbf{y}_k - \mathbf{X}_k \phi_k \right) \right\|_2^2 \right],$$

where we adjusted $(\mathbf{W}_k^*)_{TT} := (\mathbf{W}_k)_{TT}/p^*(T)$. According to our previous results, the right-hand side has to be solved for $\mu_\infty \to \infty$, and we thus set $\mu_\infty \gg 1$ to account for the limit. KernelSHAP-IQ estimates are computed by solving the *approximated* WLS objective with $\hat{\mathbf{W}}_k, \hat{\mathbf{X}}_k$ and $\hat{\mathbf{y}}_k$, containing sampled subsets $T \subseteq N$ at each row. More details and pseudocode can be found in Appendix B.

Inconsistent KernelSHAP-IQ. Our baseline method, *inconsistent KernelSHAP*, solves a single WLS problem using the KernelSHAP weights. We include all interactions up to order k via the weighting in Proposition 3.2, i.e. in total $\sum_{\ell=1}^{k} \binom{n}{\ell}$ interactions. More formally, introducing the stacked matrices $\mathbf{X}_{\leq k} := (\mathbf{X}_1, \dots, \mathbf{X}_k)$ and $\phi_{\leq k} := [\phi_1, \dots, \phi_k]^T$, this approach solves

$$\phi^* = \lim_{\mu_\infty \to \infty} \operatorname*{arg\,min}_{\phi_{\leq k}} \left\| \sqrt{\mathbf{W}_1} \left(\mathbf{y}_1 - \mathbf{X}_{\leq k} \phi_{\leq k} \right) \right\|_2^2 . \qquad (4)$$

Algorithm 1 KernelSHAP-IQ

```
Require: order $k$, sampling weights $q$, budget $b$.
$\{T_i\}_{i=1,\dots,b}, \{w_T\}_{T=T_1,\dots,T_b} \leftarrow \textsc{Sample}(q,b)$
$\mathbf{y}_1 \leftarrow [\nu(T_1), \dots, \nu(T_b)]^T$
for $\ell = 1, \dots, k$ do    ▷ iterative approximation
    for $T \in \{T_i\}$ and $|S| = \ell$ do
        $(\hat{\mathbf{X}}_\ell)_{TS} \leftarrow \lambda(|S|, |T \cap S|)$    ▷ Bernoulli weighting
        $(\hat{\mathbf{W}}_\ell^*)_{TT} \leftarrow \mu_\ell(t) \cdot w_T$    ▷ weight adjustment
    end for
    if $\ell \leq 2$ then    ▷ higher order split
        $\hat{\phi}_\ell \leftarrow \textsc{SolveWLS}(\hat{\mathbf{X}}_\ell, \hat{\mathbf{y}}_\ell, \hat{\mathbf{W}}_\ell^*)$
    else
        for $T \in \{T_i\}$ and $|S| = \ell$ do
            $(\mathbf{Q}_\ell)_{ST} \leftarrow \mathbf{1}_{T \notin \mathcal{T}_\ell} \cdot \textsc{SIIWeight}(T, S)$
            $(\hat{\mathbf{y}}_\ell^+)_T \leftarrow \mathbf{1}_{T \in \mathcal{T}_\ell} (\hat{\mathbf{y}}_\ell)_T$
            $(\hat{\mathbf{y}}_\ell^-)_T \leftarrow \mathbf{1}_{T \notin \mathcal{T}_\ell} (\hat{\mathbf{y}}_\ell)_T$
        end for
        $\hat{\phi}_\ell \leftarrow \mathbf{Q}_\ell \hat{\mathbf{y}}_\ell^- + \textsc{SolveWLS}(\hat{\mathbf{X}}_\ell, \hat{\mathbf{y}}_\ell^+, \hat{\mathbf{W}}_\ell^*)$
    end if
    $\hat{\mathbf{y}}_{\ell+1} \leftarrow \hat{\mathbf{y}}_\ell - \hat{\mathbf{X}}_\ell \cdot \hat{\phi}_\ell$    ▷ compute residuals
end for
$\hat{\Phi}_k \leftarrow \textsc{AggregateSII}(\hat{\phi}_1, \dots, \hat{\phi}_k)$    ▷ compute $k$-SII
return $k$-SII estimates $\hat{\Phi}_k$, SII estimates $\hat{\phi}_1, \dots, \hat{\phi}_k$
```

Inconsistent KernelSHAP solves the WLS problem with the approximated objective to obtain estimates for SII.

Remark 3.10. Interestingly, we observe empirically that solving Equation (4) yields the exact SVs.

Remark 3.11. Inconsistent KernelSHAP-IQ is related, but not equivalent, to k_{ADD}-SHAP (Pelegrina et al., 2023) and FSI. k_{ADD}-SHAP includes $\phi^{\text{SII}}(\emptyset)$ and a sum starting from zero in λ in $\mathbf{X}_{\leq k}$. FSI uses binary weights $\mathbf{X}_{\leq k}$, and effectively solves for k-SII Φ_k instead of SII $\phi_{\leq k}^{\text{SII}}$.

While higher-order estimates do not converge to SII, inconsistent KernelSHAP-IQ surprisingly yields high-quality estimations in low-budget settings. Utilizing our novel representations we formalize (consistent) KernelSHAP-IQ, which yields SOTA approximation *and* converges to SII.

KernelSHAP-IQ. We now introduce KernelSHAP-IQ based on the novel representation in Theorem 3.7, which ensures that KernelSHAP-IQ converges to SII for $k = 2$. Further, with Conjecture 3.9, KernelSHAP-IQ can be extended to $k > 2$, where we empirically validate its convergence. KernelSHAP-IQ once samples subsets given the sampling weights q, where the weights w account for the sampling probabilities and the number of Monte Carlo samples (line 1). As a default, we propose the KernelSHAP weights for sampling, i.e $p(T) \propto \mu_1(t)$, and apply the *border-trick* (Fumagalli et al., 2023). For further details regarding the sampling and the weights w, we refer to Appendix B.3. The game is then evaluated on all sampled subsets, stored in $\hat{\mathbf{y}}_1$, which determines the computational complexity (line 2). Then, $\hat{\nu}_\ell$ is iteratively constructed for $\ell = 1, \ldots, k$ by computing the SII estimates (line 3-19). Starting from $\ell = 1$, where KernelSHAP-IQ reduces to KernelSHAP, $\hat{\phi}_\ell$ is estimated and the residual $\hat{\mathbf{y}}_{\ell+1} := \hat{\mathbf{y}}_\ell - \mathbf{X}_\ell \cdot \hat{\phi}_\ell$ is computed for the next iteration (line 18). We repeat this process until $\ell = k$, summarized in Algorithm 1. At each step, we set the matrices for the WLS problem (line 4-7). The coefficient matrix $\hat{\mathbf{X}}_\ell$ is set for every sampled subset and every interaction of the current order ℓ (line 5). The diagonal weight matrix $\hat{\mathbf{W}}^*_\ell$ is set for every sampled subset, and contains μ_ℓ adjusted by w_T, which accounts for the sampling probabilities $p^*(T)$ and the number of Monte Carlo samples (line 6). If $\ell \leq 2$, then the WLS problem is solved directly with $\hat{\mathbf{X}}_\ell, \hat{\mathbf{y}}_\ell$, and $\hat{\mathbf{W}}^*_\ell$ (line 9). If $\ell > 2$, then Conjecture 3.9 applies, and the WLS problem is split by $\mathcal{T}_\ell$ (line 11-15). For every sampled subset and every interaction of the current order ℓ, the SII weights are assigned to $\mathbf{Q}_\ell$ (line 12), and the residuals $\hat{\mathbf{y}}_\ell$ are split by $\mathcal{T}_k$ into $\hat{\mathbf{y}}^+_\ell$ (line 13) and $\hat{\mathbf{y}}^-_\ell$ (line 14). The SII estimates $\hat{\phi}_\ell$ are then computed with the SII weights $\mathbf{Q}_\ell$ for $\hat{\mathbf{y}}^-_\ell$ and by solving the WLS problem with $\hat{\mathbf{X}}_\ell, \hat{\mathbf{y}}^+_\ell$ and $\hat{\mathbf{W}}^*_\ell$ (line 16). After reaching $\ell = k$, the SII estimates are aggregated via the non-recursive formula (cf. Appendix B.6) to k-SII values, and used for final interpretation (line 20-21).

4. Experiments

We conduct multiple experiments to compare KernelSHAP-IQ with existing baseline methods for estimating SII and k-SII values. For each method, we assess estimation quality with *mean-squared error* (MSE; lower is better) and *precision at ten* (Prec@10; higher is better) compared to ground truth (GT) SII and k-SII. We compute confidence bands with the standard error of the mean (SEM). Prec@10 measures the accuracy of correctly identifying the ten highest interaction scores in terms of absolute values. The GT values are calculated once via brute force by evaluating 2^n coalitions for each game of n players (i.e. features).

Baselines. We compare KernelSHAP-IQ and inconsistent KernelSHAP-IQ with all available baseline algorithms; permutation sampling (Castro et al., 2009; Tsai et al., 2023), SHAP-IQ (Fumagalli et al., 2023), and SVARM-IQ (Kolpaczki et al., 2024b), as shown in Table 1.

Table 2. Summary of the benchmark datasets and models used.

ID	Model	Removal Strategy	n	$\mathcal{Y}$
SOUM	Synthetic	–	20, 40	$[0, 1]$
LM	DistilBert	Token Removal	14	$[-1, 1]$
CH	Neural Net	Mean	8	$\mathbb{R}$
BR	XGBoost	Mean/Mode	12	$\mathbb{R}$
ViT	ViT-32-384	Token Removal	16	$[0, 1]$
CNN	ResNet18	Superpixel	14	$[0, 1]$
AC	RF	Mean/Mode	14	$[0, 1]$

Benchmark Datasets and Models. Based on recent work by Fumagalli et al. (2023); Kolpaczki et al. (2024b); Tsai et al. (2023), we create different benchmark scenarios[2]. Table 2 summarizes the scenarios and corresponding removal approaches after Covert et al. (2021). For a detailed description regarding the experimental setup and model training we refer to Appendices C and D including a runtime analysis. First, we create synthetic *sum of unanimity models* (SOUMs), also known as sum of unanimity games and induced subgraph game (Deng & Papadimitriou, 1994), where GT SII can be computed, cf. Appendix B.8. Second, we compute feature (i.e. token) interaction values for a sentiment analysis *language model* (LM). The LM is a fine-tuned version of `DistilBert` (Sanh et al., 2019) on movie review excerpts from the IMDB dataset (Maas et al., 2011). Third and forth, we explain local instances of the *bike rental* (BR) (Fanaee-T & Gama, 2014) and *california housing* (CH) (Kelley Pace & Barry, 1997) regression datasets. On BR we train an XGBoost regressor and on CH a small neural network. The target variables are logarithmized. Fifth and sixth, we explain a `ResNet18` *convolutional neural network* (CNN) (He et al., 2016) and a *vision transformer* (ViT) image classifiers, which were pre-trained on ImageNet (Deng et al., 2009). The ViT considers patches of 32×32 pixels. In the CNN, individual pixels are grouped together into super-pixels. Seventh, we compute feature interactions for instances of the *adult census* (AC) (Kohavi, 1996) classification dataset and a random forest (RF) classifier.

Approximation Quality of SII and k-SII. Figure 4 depicts the approximation quality of KernelSHAP-IQ and inconsistent KernelSHAP-IQ compared to the baseline estimators on a selection of benchmark tasks. Further results are provided in Appendix D. KernelSHAP-IQ and inconsistent KernelSHAP-IQ clearly outperform all three sampling-based baselines on both regression benchmarks (BR and CH). Inconsistent KernelSHAP-IQ achieves high estimation qualities in low-budget scenarios. This trend also materializes for second order SII values, in particular in high-player tasks (e.g. SOUM in Figure 4). Yet, unlike KernelSHAP-IQ, the inconsistent version does not converge to the GT SII values. Interestingly, we observe that the inconsistent version

[2] `https://github.com/FFmgll/KernelSHAP_IQ_Supplementary_Material`

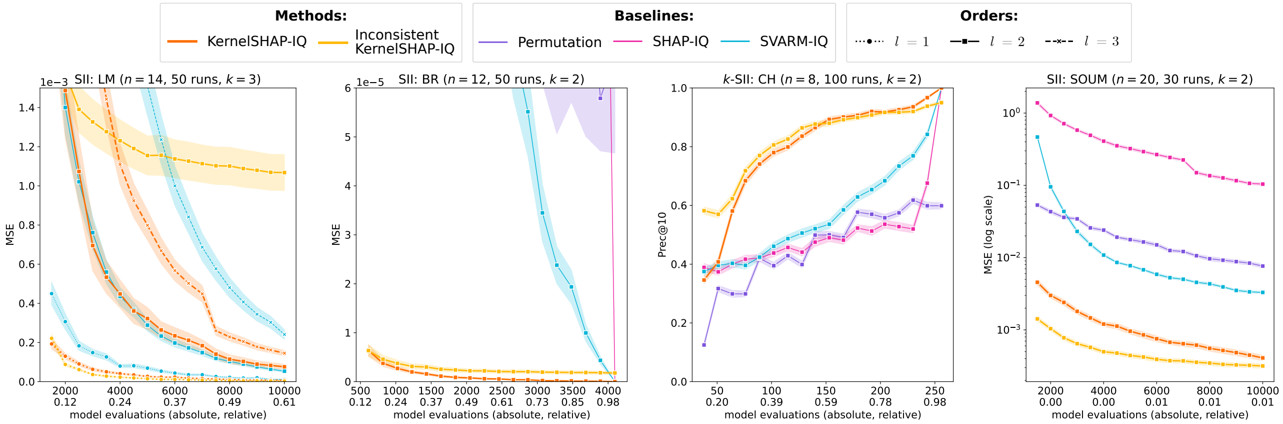

Figure 4. Approximation quality of KernelSHAP-IQ (orange) and inconsistent KernelSHAP-IQ (yellow) compared to the permutation sampling (purple), SHAP-IQ (pink) and SVARM-IQ (blue) baselines for estimating SII values for the LM (left; $n = 14, l \in \{1, 2, 3\}$) the *bike rental* dataset (center left, $n = 12, l = 2$), the *california housing* dataset (center right; $n = 8, l = 2$), and the SOUM (right; $n = 20, l = 2$). The shaded bands represent the standard error of the mean (SEM).

does converge to the GT SV independent of the approximation order $k \geq 1$. Notably, in benchmark tasks like the LM, KernelSHAP-IQ and SVARM-IQ rapidly outperform all baselines as well as inconsistent KernelSHAP-IQ. Both SVARM-IQ and KernelSHAP-IQ achieve SOTA results for the LM task and second order SII values. By relying on Conjecture 3.9, KernelSHAP-IQ yields high-quality estimations for SII of higher orders ($k > 2$) and outperforms SVARM-IQ for order 3 on the LM.

Example Use Case: Feature Attribution. As illustrated in Figures 1 and 2, 2-SII values can be used to enhance current feature attribution techniques. Similar to Fumagalli et al. (2023); Sundararajan et al. (2020); Tsai et al. (2023), we show that feature interactions are relevant for understanding intricate LMs. In Figure 1, 2-SII scores reveal that the predicted positive sentiment largely stems from the interaction of the two words “never” and “forget”, while “forget” individually points towards a negative sentiment. In the CH example illustrated in Figure 2, the *longit.* and *latit.* features are both contributing positively to the prediction considering 1-SII (SV). The 2-SII explanation, however, reveals that the positive contribution of *latit.* vanishes (very low *latit.* score) and can be mostly attributed to the *latit. x longit.* interaction. Hence, the *exact location* of the property is meaningful.

5. Limitations

We linked the SII to a solution of a WSL optimization problem. As previously discussed, our theoretical results extend to higher orders, provided that $2k \leq n$, although we were unable to give a rigorous proof. We expect that other proof techniques are required to further understand these coherences, which could resolve improper weighting in Conjecture 3.9. In practice, due to iterative computation of KernelSHAP-IQ, higher-order estimates are negatively affected by previous low-quality estimates, where interpretation may be flawed. Lastly, the exponentially growing number of interactions requires human-centered post processing, and we give modest suggestions for visualizations.

6. Conclusion and Future Work

In this work, we clarified the link between SII, an axiomatic interaction index, and k-SII, an aggregation of SII that yields a k-additive interaction index used for interpretability. We demonstrated that the approximation of any game induced by k-SII is iteratively constructed via SII. Similar to the SV, we then established that SII of order k is represented as the solution to a WLS problem, where SII yields an *optimal* k-additive approximation. We rigorously prove our results for the SV and pairwise SII and give empirically validated conjectures for higher orders. Consequently, we introduce KernelSHAP-IQ, a direct extension of KernelSHAP, which efficiently approximates SII for higher orders and yields SOTA performance. We apply KernelSHAP-IQ for local interpretability and demonstrate benefits of enriching feature attributions with interactions.

In future work, we suspect that a rigorous proof of our conjectures would reveal novel techniques and further insights. Another interesting link are orthogonal projections, where we suspect a link of pairwise SII to the *Shapley residual* (Kumar et al., 2021). Our novel representation of SII opens up additional possibilities, including the investigation of *amortized* Shapley interactions through methods akin to FastSHAP (Jethani et al., 2022). Apart from local interpretability with feature interactions, KernelSHAP-IQ is applicable in any game-theoretic setting, which includes global interpretability or data valuation.

Acknowledgment

We gratefully thank the anonymous reviewer for their valuable feedback for improving this work. Fabian Fumagalli and Maximilian Muschalik gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824. Patrick Kolpaczki 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.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically the field of Explainable Artificial Intelligence. There are many potential societal consequences of our work. Our work can positively impact Machine Learning adoption and potentially reveal biases or unwanted behavior in Machine Learning systems. However, it may also enable explainability-based white-washing in organizations, firms, or policing.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120.

Banzhaf III, J. F. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317, 1964.

Bordt, S. and von Luxburg, U. From Shapley Values to Generalized Additive Models and back. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 709–745. PMLR, 2023.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. doi: 10.1016/j.cor.2008.04.004.

Charnes, A., Golany, B., Keane, M., and Rousseau, J. *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, volume 11, pp. 123–133. Springer Netherlands, 1988. doi: 10.1007/978-94-009-3677-5_7.

Chen, H., Covert, I., Lundberg, S., et al. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5:590–601, 2023. doi: 10.1038/s42256-023-00657-x.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2016)*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Covert, I. and Lee, S. 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*, pp. 3457–3465. PMLR, 2021.

Covert, I., Lundberg, S. M., and Lee, S. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. doi: 10.5555/3546258.3546467.

Deng, H., Zou, N., Du, M., Chen, W., Feng, G., Yang, Z., Li, Z., and Zhang, Q. Unifying Fourteen Post-hoc Attribution Methods with Taylor Interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2024. doi: 10.1109/TPAMI.2024.3358410.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 2009)*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848.

Deng, X. and Papadimitriou, C. H. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994. doi: 10.1287/moor.19.2.257.

Fanaee-T, H. and Gama, J. Event Labeling Combining Ensemble Detectors and Background Knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014. doi: 10.1007/s13748-013-0040-3.

Feurer, M., van Rijn, J. N., Kadra, A., Gijsbers, P., Mallik, N., Ravi, S., Mueller, A., Vanschoren, J., and Hutter, F. OpenML-Python: an extensible Python API for OpenML. *CoRR*, abs/1911.02490, 2020.

Fujimoto, K., Kojadinovic, I., and Marichal, J. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006. doi: 10.1016/j.geb.2005.03.002.

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. E. SHAP-IQ: Unified approximation of any-order shapley interactions. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

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 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019.

Grabisch, M. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2): 167–189, 1997. doi: 10.1016/S0165-0114(97)00168-1.

Grabisch, M. *Set Functions, Games and Capacities in Decision Making*, volume 46. Springer International Publishing Switzerland, 2016. ISBN 978-3-319-30690-2. doi: 10.1007/978-3-319-30690-2.

Grabisch, M. and Roubens, M. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4): 547–565, 1999. doi: 10.1007/s001820050125.

Grabisch, M., Marichal, J., and Roubens, M. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000. doi: 10.1287/moor.25.2.157.12225.

Hammer, P. L. and Holzman, R. Approximations of pseudo-boolean functions; applications to game theory. *ZOR Mathematical Methods of Operations Research*, 36(1): 3–21, 1992. doi: 10.1007/BF01541028.

Harris, C., Pymar, R., and Rowat, C. Joint shapley values: a measure of joint feature importance. In *The Tenth International Conference on Learning Representations, (ICLR 2022)*. OpenReview.net, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR 2016)*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Herbinger, J., Bischl, B., and Casalicchio, G. Decomposing Global Feature Effects Based on Feature Interactions. *CoRR*, abs/2306.00541, 2023.

Herren, A. and Hahn, P. R. Statistical aspects of shap: Functional anova for model interpretation. *CoRR*, abs/2208.09970, 2022.

Hiabu, M., Meyer, J. T., and Wright, M. N. Unifying local and global model explanations by functional decomposition of low dimensional structures. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7040–7060. PMLR, 2023.

Hooker, G. Discovering additive structure in black box functions. In Kim, W., Kohavi, R., Gehrke, J., and DuMouchel, W. (eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2004)*, pp. 575–580. ACM, 2004. doi: 10.1145/1014052.1014122.

Hooker, G. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16 (3):709–732, 2007. doi: 10.1198/106186007X237892.

Janizek, J. D., Sturmfels, P., and Lee, S. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021.

Jethani, N., Sudarshan, M., Covert, I. C., Lee, S., and Ranganath, R. FastSHAP: Real-Time Shapley Value Estimation. In *The Tenth International Conference on Learning Representations (ICLR 2022)*. OpenReview.net, 2022.

Kelley Pace, R. and Barry, R. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: https://doi.org/10.1016/S0167-7152(96)00140-X.

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 1996)*, pp. 202–207, 1996.

Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. Approximating the Shapley Value without Marginal Contributions. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, (AAAI 2024)*, pp. 13246–13255. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I12.29225.

Kolpaczki, P., Muschalik, M., Fumagalli, F., Hammer, B., and Hüllermeier, E. SVARM-IQ: Efficient approximation of any-order Shapley interactions through stratification. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, (AISTATS 2024)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, 2024b.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. A. 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*, pp. 26598–26608, 2021.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. 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*, pp. 5491–5500. PMLR, 2020.

Lengerich, B. J., Tan, S., Chang, C., Hooker, G., and Caruana, R. 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*, pp. 2402–2412. PMLR, 2020.

Lhoest, Q., Villanova del Moral, A., von Platen, P., Wolf, T., Šaško, M., Jernite, Y., Thakur, A., Tunstall, L., Patil, S., Drame, M., Chaumond, J., Plu, J., Davison, J., Brandeis, S., Sanh, V., Le Scao, T., Canwen Xu, K., Patry, N., Liu, S., McMillan-Major, A., Schmid, P., Gugger, S., Raw, N., Lesage, S., Lozhkov, A., Carrigan, M., Matussière, T., von Werra, L., Debut, L., Bekman, S., and Delangue, C. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, (EMNLP 2021)*, pp. 175–184. Association for Computational Linguistics, 2021.

Lundberg, S. M. and Lee, S. 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*, pp. 4765–4774, 2017.

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. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (HLT 2011)*, pp. 142–150. Association for Computational Linguistics, 2011.

Marichal, J.-L. and Roubens, M. *The Chaining Interaction Index among Players in Cooperative Games*, pp. 69–85. Springer Netherlands, 1999. doi: 10.1007/978-94-017-0647-6_5.

Murdoch, W. J., Liu, P. J., and Yu, B. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *6th International Conference on Learning Representations, (ICLR 2018)*, 2018.

Muschalik, M., Fumagalli, F., Hammer, B., and Hüllermeier, E. Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, (AAAI 2024)*, pp. 14388–14396. AAAI Press, 2024. doi: 10.1609/AAAI.V38I13.29352.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, (NeurIPS 2019)*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

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, 2023. doi: 10.1016/J.ARTINT.2023.104014.

Ruiz, L. M., Valenciano, F., and Zarzuelo, J. M. The family of least square values for transferable utility games. *Games and Economic Behavior*, 24(1):109–130, 1998. doi: https://doi.org/10.1006/game.1997.0622.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

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.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES 2020)*, pp. 180–186. ACM, 2020. doi: 10.1145/3375627.3375830.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. 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*, pp. 9259–9268. PMLR, 2020.

Tsai, C., Yeh, C., and Ravikumar, P. Faith-Shap: The Faithful Shapley Interaction Index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Tsang, M., Cheng, D., and Liu, Y. Detecting statistical interactions from neural network weights. In *6th International Conference on Learning Representations, (ICLR 2018)*, 2018.

Tsang, M., Cheng, D., Liu, H., Feng, X., Zhou, E., and Liu, Y. Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. In *8th International Conference on Learning Representations, (ICLR 2020)*, 2020a.

Tsang, M., Rambhatla, S., and Liu, Y. How does This Interaction Affect Me? Interpretable Attribution for Feature Interactions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, pp. 6147–6159, 2020b.

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., and Biernacka, J. M. SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinformatics*, 13:164, 2012. doi: 10.1186/1471-2105-13-164.

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. 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)*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Wright, M. N., Ziegler, A., and König, I. R. Do little interactions get lost in dark random forests? *BMC Bioinform.*, 17:145, 2016. doi: 10.1186/s12859-016-0995-8.

Yu, P., Bifet, A., Read, J., and Xu, C. Linear tree shap. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, (NeurIPS 2022)*, 2022.

Zern, A., Broelemann, K., and Kasneci, G. Interventional SHAP values and interaction values for piecewise linear regression trees. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, (AAAI 2023)*, pp. 11164–11173. AAAI Press, 2023. doi: 10.1609/AAAI.V37I9.26322.

Zhang, H., Xie, Y., Zheng, L., Zhang, D., and Zhang, Q. Interpreting Multivariate Shapley Interactions in DNNs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI 2021)*, pp. 10877–10886. AAAI Press, 2021.

Organisation of the Supplementary Material

The supplementary material is organized as follows. Appendix A contains all proofs. Appendix B contains algorithmic details and ground truth computation procedures. Appendix C contains details regarding the experimental setup and reproducibility. Lastly, Appendix D contains additional experimental results.

A Proofs **15**
A.1 Proof of Proposition 3.2 . . . 15
A.2 Proof of Corollary 3.3 . . . 15
A.3 Proof of Corollary 3.4 . . . 15
A.4 Proof of Theorem 3.6 . . . 16
A.5 Proof of Theorem 3.7 . . . 17

B Algorithms and Ground Truth Calculations **23**
B.1 KernelSHAP-IQ Default Parameters . . . 23
B.2 Inconsistent KernelSHAP-IQ . . . 23
B.3 Sampling Algorithm . . . 23
B.4 Solving WLS . . . 24
B.5 Subset Weights SII . . . 25
B.6 Aggregate SII to k-SII . . . 25
B.7 Baseline Methods: SHAP-IQ, Permutation Sampling and SVARM-IQ . . . 26
B.8 Analytic Solution for SII of SOUMs . . . 26
B.9 Intuition about the KernelSHAP-IQ Weights . . . 27

C Experimental Setup and Reproducibility **28**
C.1 Model, Datasets and Task Descriptions . . . 28
C.2 Computational Effort . . . 29

D Additional Empirical Results **30**
D.1 Runtime Analysis . . . 30
D.2 Validations of Higher-Order Conjecture . . . 30
D.3 Additional Approximation Results . . . 31

A. Proofs

A.1. Proof of Proposition 3.2

Proof. By Definition 3.1, we have $\hat{\nu}_k(T) := \sum_{S \subseteq T}^{1 \leq |S| \leq k} \Phi_k(S)$. According to Appendix A.2 in (Bordt & von Luxburg, 2023), $\Phi_k(S)$ is explicitly computed as

$$\Phi_k(S) = \sum_{S \subseteq \tilde{S} \subseteq N}^{|\tilde{S}| \leq k} B_{|\tilde{S}|-|S|} \phi^{\text{SII}}(\tilde{S}) \text{ for } 1 \leq |S| \leq k.$$

Hence, it follows by re-arranging terms

$$\begin{aligned}
\hat{\nu}_k(T) &= \sum_{S \subseteq T}^{1 \leq |S| \leq k} \Phi_k(S) = \sum_{S \subseteq T}^{1 \leq |S| \leq k} \sum_{S \subseteq \tilde{S} \subseteq N}^{|\tilde{S}| \leq k} B_{|\tilde{S}|-|S|} \phi^{\text{SII}}(\tilde{S}) = \sum_{S \subseteq T}^{1 \leq |S|} \sum_{\tilde{S} \subseteq N}^{|\tilde{S}| \leq k} B_{|\tilde{S}|-|S|} \phi^{\text{SII}}(\tilde{S}) \mathbf{1}(S \subseteq \tilde{S}) \\
&= \sum_{\tilde{S} \subseteq N}^{|\tilde{S}| \leq k} \phi^{\text{SII}}(\tilde{S}) \sum_{S \subseteq T}^{1 \leq |S|} B_{|\tilde{S}|-|S|} \mathbf{1}(S \subseteq \tilde{S}) = \sum_{\tilde{S} \subseteq N}^{|\tilde{S}| \leq k} \phi^{\text{SII}}(\tilde{S}) \sum_{r=1}^{|\tilde{S}|} B_{|\tilde{S}|-r} \sum_{S \subseteq N}^{|S|=r} \mathbf{1}(S \subseteq \tilde{S} \cap T) \\
&= \sum_{\tilde{S} \subseteq N}^{|\tilde{S}| \leq k} \phi^{\text{SII}}(\tilde{S}) \sum_{r=1}^{|\tilde{S} \cap T|} B_{|\tilde{S}|-r} \binom{|\tilde{S} \cap T|}{r} = \sum_{\tilde{S} \subseteq N}^{|\tilde{S}| \leq k} \phi^{\text{SII}}(\tilde{S}) \lambda(|\tilde{S}|, |\tilde{S} \cap T|),
\end{aligned}$$

where we introduced $\binom{0}{r} := 0$ for $r > 0$ and $\lambda(k, \ell) := \sum_{r=1}^{\ell} \binom{\ell}{r} B_{k-r}$. The result follows then immediately by observing that the terms do not depend on k and separating above sum into terms of order k and order less than k as

$$\hat{\nu}_k(T) = \hat{\nu}_{k-1}(T) + \sum_{\tilde{S} \subseteq N}^{|\tilde{S}|=k} \phi^{\text{SII}}(\tilde{S}) \lambda(|\tilde{S}|, |\tilde{S} \cap T|).$$

□

A.2. Proof of Corollary 3.3

Proof. From Proposition 3.2 it suffices to compute $\lambda(k, \ell) := \sum_{r=1}^{\ell} \binom{\ell}{r} B_{k-r}$. Clearly, $\lambda(1, 1) = B_0 = 1$, which yields $\hat{\nu}_1(T) = \sum_{i \in T} \phi^{\text{SII}}(i) = \sum_{i \in T} \phi^{\text{SV}}(i)$. For $k > 1$, we have by the symmetry of the binomial coefficient $\lambda(k, k) = \sum_{r=1}^{k} \binom{k}{r} B_{k-r} = \sum_{r=0}^{k-1} \binom{k}{r} B_k = 0$, which is the defining property of the Bernoulli numbers. Hence, for $k = 2$, $\lambda(2, 0) = \lambda(2, 2) = 0$. Lastly, $\lambda(2, 1) = B_1 = -1/2$, which yields

$$\hat{\nu}_2(T) = \hat{\nu}_1(T) - \frac{1}{2} \sum_{ij \subseteq N}^{|ij \cap T|=1} \phi^{\text{SII}}(ij).$$

□

A.3. Proof of Corollary 3.4

Proof. This proof follows immediately from the fact that $\hat{\nu}_n(T)$ is constructed from k-SII values Φ_k, which are the Möbius transform for $k = n$, cf. Theorem 4 by Bordt & von Luxburg (2023). A defining property of the Möbius transform (Grabisch, 2016) is

$$\nu(T) = \sum_{S \subseteq T} \Phi_n(S),$$

and thus, by Definition 3.1, $\nu(T) = \hat{\nu}_n(T)$. □

A.4. Proof of Theorem 3.6

Proof. The solution of the optimization problem is explicitly given as

$$\phi^* = (\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{W}_1 \cdot \mathbf{y}_1. \tag{5}$$

Our goal is show that for $\mu_\infty \to \infty$ we obtain the correct weight of the SV for $\left((\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{W}_1\right)_{iT}$. We will show that those weights are equal to the representation given in Theorem 4.4 by Fumagalli et al. (2023) as

$$\phi^{\text{SV}}(i) = \frac{1}{n}\nu(N) + \sum_{T \subseteq N}^{1 \leq |T| \leq n-1} \nu(T) \frac{\mu_1(t)}{n-1} \left[\mathbf{1}(i \in T) - \frac{t}{n}\right].$$

Note that the values $\nu(T)$ are encoded in $\mathbf{y}_1$. The proof is structured as follows:

- Find structure of $\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1$.
- Find inverse $(\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1}$.
- Compute $\lim_{\mu_\infty \to \infty} (\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{W}_1$
  - for T with finite weight, i.e. $1 \leq |T| \leq n-1$.
  - for T with infinite weight $\mu_1(t) = \mu_\infty$, i.e. $T \in \{\emptyset, N\}$.

We proceed by computing the terms in ϕ^* separately. First, with $(\mathbf{X}_1)_{Ti} = \mathbf{1}(i \in T)$

$$(\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)_{i,j} = \sum_{T \subseteq N} \mu_1(|T|) \mathbf{1}(i \in T) \mathbf{1}(j \in T) = \mathbf{1}(i = j) \underbrace{\sum_{t=1}^{n} \mu_1(t) \binom{n-1}{t-1}}_{=:\mu_{1,1}} + \mathbf{1}(i \neq j) \underbrace{\sum_{t=2}^{n} \mu_1(t) \binom{n-2}{t-2}}_{=:\mu_{1,0}}.$$

Introducing a matrix $(\mathbf{J})_{i,j} \equiv 1$ of all ones and the identity $\mathbf{I}$, we have $(\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1) = \mu_{1,0}\mathbf{J} + (\mu_{1,1} - \mu_{1,0})\mathbf{I}$. Due to this simplistic structure, we can compute the inverse explicitly using the following Lemma.

Lemma A.1 (Fumagalli et al. (2023))**.** *Let $\eta_0, \eta_1 > 0$ with $\eta_0 \neq \eta_1$. Then, $(\eta_0 \mathbf{J} + (\eta_1 - \eta_0)\mathbf{I})^{-1} = \tilde{\eta}_0 \mathbf{J} + (\tilde{\eta}_1 - \tilde{\eta}_0)\mathbf{I}$ with*

$$\tilde{\eta}_0 = \frac{-\eta_0}{(\eta_1 - \eta_0)(\eta_1 + (n-1)\eta_0)} \qquad \tilde{\eta}_1 = \frac{\eta_1 + (n-2)\eta_0}{(\eta_1 - \eta_0)(\eta_1 + (n-1)\eta_0)}.$$

Hence,

$$(\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1} = (\mu_{1,0}\mathbf{J} + (\mu_{1,1} - \mu_{1,0})\mathbf{I})^{-1} = \tilde{\mu}_{1,0}\mathbf{J} + (\tilde{\mu}_{1,1} - \tilde{\mu}_{1,0})\mathbf{I}.$$

By definition of μ_1, we have

$$\mu_{1,1} = \mu_\infty + \sum_{t=1}^{n-1} \binom{n-2}{t-1}^{-1} \binom{n-1}{t-1} = \mu_\infty + \sum_{t=1}^{n-1} \frac{n-1}{n-t} = \mu_\infty + (n-1)h_{n-1} = \mu_\infty + (n-1)h_{n-2} + 1,$$

where $h_n := \sum_{k=1}^{n} 1/k$ is the harmonic number. Further,

$$\mu_{1,0} = \mu_\infty + \sum_{t=2}^{n-1} \binom{n-2}{t-1}^{-1} \binom{n-2}{t-2} = \mu_\infty + \sum_{t=2}^{n-1} \frac{t-1}{n-t} = \mu_\infty + \sum_{t=1}^{n-2} \frac{n-t-1}{t} = \mu_\infty + (n-1)h_{n-2} - (n-2).$$

Hence, $\mu_{1,1} - \mu_{1,0} = n-1$. Note that if $\mu_\infty \to \infty$, then $\tilde{\mu}_{1,0} \overset{\mu_\infty \to \infty}{\longrightarrow} \frac{1}{n(n-1)}$ and $\tilde{\mu}_{1,1} \overset{\mu_\infty \to \infty}{\longrightarrow} \frac{1}{n}$, which proves the special case for $k=1$ of Conjecture 3.8. We are now able to compute the weights

$$(\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{W}_1 = \tilde{\mu}_{1,0}\mathbf{J} \cdot \mathbf{X}_1^T \mathbf{W}_1 + (\tilde{\mu}_{1,1} - \tilde{\mu}_{1,0})\mathbf{I} \cdot \mathbf{X}_1^T \mathbf{W}_1.$$

By Lemma A.1, $\tilde{\mu}_{1,1} - \tilde{\mu}_{1,0} = \frac{1}{\mu_{1,1}-\mu_{1,0}} = \frac{1}{n-1}$. With $(\mathbf{J}\mathbf{X}_1^T\mathbf{W}_1)_{ST} = |T|\mu_1(t)$ it follows

$$\begin{aligned}\left(\tilde{\mu}_{1,0}\mathbf{J}\cdot\mathbf{X}_1^T\mathbf{W}_1 + (\tilde{\mu}_{1,1}-\tilde{\mu}_{1,0})\mathbf{I}\cdot\mathbf{X}_1^T\mathbf{W}_1\right)_{iT} &= \tilde{\mu}_{1,0}|T|\mu_1(t) + \frac{1}{n-1}\mathbf{1}(i\in T)\mu_1(t)\\ &= \frac{1}{n-1}\mu_1(t)\left(\frac{-\mu_{1,0}|T| + \mathbf{1}(i\in T)(\mu_{1,1}+(n-1)\mu_{1,0})}{\mu_{1,1}+(n-1)\mu_{1,0}}\right).\end{aligned}$$

Now, for T with $\mu_1(t) \neq \mu_\infty$, i.e. $1 \leq |T| \leq n-1$, we have that $\mu_1(t) \overset{\mu_\infty\to\infty}{\longrightarrow} \mu_1(t)$ and thus compute

$$\left((\mathbf{X}_1^T\mathbf{W}_1\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{W}_1\right)_{iT} \overset{\mu_\infty\to\infty}{\longrightarrow} \frac{1}{n-1}\mu_1(t)\left(\mathbf{1}(i\in T) - \frac{|T|}{n}\right),$$

where we used that μ_∞ appears in $\mu_{1,0}$ and $\mu_{1,1}$. Clearly, this directly yields the weight in the representation of Theorem 4.4 in (Fumagalli et al., 2023). For the cases, where $\mu_1(t) = \mu_\infty$, i.e. $T \in \{\emptyset, N\}$, we have zero weight for $\emptyset$, as $|T| = 0$ and $\mathbf{1}(i \in T) = 0$ for all $i \in N$. For $T = N$, we have $|T| = n$ and $\mathbf{1}(i\in T) = 1$ for all $i \in N$. Hence,

$$\left((\mathbf{X}_1^T\mathbf{W}_1\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{W}_1\right)_{iN} = \frac{1}{n-1}\mu_1(t)\left(\frac{\mu_{1,1}-\mu_{1,0}}{\mu_{1,1}+(n-1)\mu_{1,0}}\right) = \mu_1(t)\left(\frac{1}{\mu_{1,1}+(n-1)\mu_{1,0}}\right) \overset{\mu_\infty\to\infty}{\longrightarrow} \frac{1}{n},$$

which is again the weight in Theorem 4.4 in (Fumagalli et al., 2023). □

A.5. Proof of Theorem 3.7

Proof. Again, the solution of the optimization problem is explicitly given as

$$\phi^* = (\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\cdot\mathbf{y}_2. \tag{6}$$

We show that for $\mu_\infty\to\infty$, we obtain the correct weight of the SII in $\left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{ST}$ with $S = ij$. We will show that those weights are equal to the representation given in Theorem 4.1 in (Fumagalli et al., 2023) with $|S| = 2$ as

$$\phi^{\text{SII}}(S) = \sum_{T\subseteq N}\nu(T)(-1)^{|S|-|T\cap S|}\frac{1}{n-|S|+1}\binom{n-|S|}{|T|-|T\cap S|}^{-1} = \frac{1}{n-1}\sum_{T\subseteq N}\nu(T)(-1)^{|T\cap S|}\binom{n-2}{|T|-|T\cap S|}^{-1}. \tag{7}$$

Note that for $T\in\{\emptyset, N\}$, we have $(\mathbf{y}_2)_T = \nu(T) - \hat{\nu}_1(T) = \nu(T) - \sum_{i\in T}\phi^{\text{SV}}(i) = 0$ and thus we do not have to consider these cases. The proof is again structured as follows:

- Find structure of $\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2$.
- Find conditions for inverse $(\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}$.
- Compute $\lim_{\mu_\infty\to\infty}(\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2$
  - for T with finite weight, i.e. $2 \leq |T| \leq n-2$.
  - for T with infinite weight $\mu_2(t) = \mu_\infty$, i.e. $|T| = 1$ or $|T| = n-1$.

Recall from Corollary 3.3 that $\lambda(2,0) = \lambda(2,2) = 0$ and $\lambda(2,1) = -1/2$. Hence,

$$\left(\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2\right)_{S_1S_2} = \frac{1}{4}\sum_{T\subseteq N}\mu_2(t)\mathbf{1}(|T\cap S_1| = 1)\mathbf{1}(|T\cap S_2| = 1) =: \mu_{2,|S_1\cap S_2|}.$$

For each T, the element may either be from $S_1 \cap S_2$ or from $T \setminus S_1$ and $T \setminus S_2$, and thus

$$\begin{aligned}
\mu_{2,0} &= \frac{1}{4}\sum_{t=2}^{n-2}\binom{n-4}{t-2}^{-1}\binom{n-4}{t-2}4 = n-3 \\
\mu_{2,1} &= \frac{1}{2}\mu_\infty + \frac{1}{4}\sum_{t=2}^{n-2}\binom{n-4}{t-2}^{-1}\left[\binom{n-3}{t-1}+\binom{n-3}{t-2}\right] = \frac{1}{2}\mu_\infty + \frac{1}{4}\sum_{t=2}^{n-2}\frac{(n-2)(n-3)}{(n-t-1)(t-1)} \\
&= \frac{1}{2}\mu_\infty + \frac{n-3}{4}\sum_{t=2}^{n-2}\left(\frac{1}{t-1}+\frac{1}{n-t-1}\right) = \frac{1}{2}\mu_\infty + \frac{n-3}{2}h_{n-3}. \\
\mu_{2,2} &= \mu_\infty + \frac{1}{4}\sum_{t=2}^{n-2}\binom{n-4}{t-2}^{-1}\binom{n-2}{t-1}2 = \mu_\infty + (n-3)h_{n-3} = 2\mu_{2,1},
\end{aligned}$$

where for $|T| = 1$ or $|T| = n-1$ with weight μ_∞ no combination is found for $|S_1 \cap S_2| = \emptyset$, two for $|S_1 \cap S_2| = 1$, i.e. $T = S_1 \cap S_2$ and $T = N \setminus (S_1 \cap S_2)$ and four for $S_1 = S_2$, i.e. $T = i$ and $T = N \setminus i$ with $i \in S_1$. Due to this simplistic structure of $\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2$, we now introduce three square matrices indexed with all subsets of size 2, i.e. $\binom{n}{2}$ many. We introduce the matrix with all ones, $\mathbf{J}$, the identity $\mathbf{I}$ and the *intersection matrix* $\mathbf{Q} := \mathbf{1}(S_1 \cap S_2 = \emptyset)$. Using these matrices, we can rewrite

$$\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2 = \mu_{2,1}\mathbf{J} + (\mu_{2,0}-\mu_{2,1})\mathbf{Q} + (\mu_{2,2}-\mu_{2,1})\mathbf{I}.$$

We now first prove the following lemma.

Lemma A.2. *Let $\eta_0, \eta_1, \eta_2 > 0$, $q_k := \binom{n-4+k}{2}$ for $k = 0, \dots, 2$ and let square matrices $\mathbf{J}, \mathbf{Q}, \mathbf{I}$ indexed by all subsets of size 2 of N. Then, $(\eta_1\mathbf{J} + (\eta_0-\eta_1)\mathbf{Q} + (\eta_2-\eta_1)\mathbf{I})^{-1} = \tilde\eta_1\mathbf{J} + (\tilde\eta_0-\tilde\eta_1)\mathbf{Q} + (\tilde\eta_2-\tilde\eta_1)\mathbf{I}$, where*

$$\begin{aligned}
\tilde\eta_2 - \tilde\eta_1 &= \frac{(\eta_2-\eta_1)+(q_0-q_1)(\eta_0-\eta_1)}{(\eta_2-\eta_1)^2+(q_0-q_1)(\eta_2-\eta_1)(\eta_0-\eta_1)-(q_2-q_1)(\eta_0-\eta_1)^2}, \\
\tilde\eta_0 - \tilde\eta_1 &= -\frac{(\eta_0-\eta_1)(\tilde\eta_2-\tilde\eta_1)}{\eta_2-\eta_1+(q_0-q_1)(\eta_0-\eta_1)} \\
\tilde\eta_1 &= -\frac{(q_2\eta_1+q_1(\eta_0-\eta_1))(\tilde\eta_0-\tilde\eta_1)+\eta_1(\tilde\eta_2-\tilde\eta_1)}{(\binom{n}{2}-1-q_2)\eta_1+q_2\eta_0+\eta_2}
\end{aligned}$$

provided that the inverse exists and all denominators are unequal zero.

Remark A.3. Note that this system of equations can be directly solved by computing the first equation. Then, inserting this solution into the second equation and solving it explicitly. Finally inserting both results into the third equation to compute $\tilde\eta_1$, and consequently, $\tilde\eta_0$ and $\tilde\eta_2$. However, we do not need this explicit structure and will rely on the above conditions to prove our result.

Proof of Lemma A.2. We explicitly compute the product

$$(\eta_1\mathbf{J} + (\eta_0-\eta_1)\mathbf{Q} + (\eta_2-\eta_1)\mathbf{I})\left(\tilde\eta_1\mathbf{J} + (\tilde\eta_0-\tilde\eta_1)\mathbf{Q} + (\tilde\eta_2-\tilde\eta_1)\mathbf{I}\right).$$

In this computation the products $\mathbf{J}^2 = \binom{n}{2}\mathbf{J}$, $\mathbf{JI} = \mathbf{IJ} = \mathbf{J}$, $\mathbf{I}^2 = \mathbf{I}$, $\mathbf{QI} = \mathbf{IQ} = \mathbf{Q}$ are trivial to compute. Additionally,

$$(\mathbf{Q}^2)_{S_1S_2} = \sum_{S\subseteq N}^{|S|=2}\mathbf{1}(S_1\cap S=\emptyset)\mathbf{1}(S\cap S_2=\emptyset) = \sum_{S\subseteq N}^{|S|=2}\mathbf{1}((S_1\cup S_2)\cap S=\emptyset) = \binom{n-|S_1\cup S_2|}{2} = q_{|S_1\cap S_2|},$$

since $|S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2| = 4 - |S_1 \cap S_2|$ and $\mathbf{JQ} = \mathbf{QJ} = q_2\mathbf{J}$. Hence, we can write $\mathbf{Q}^2 =$

$q_1\mathbf{J} + (q_0 - q_1)\mathbf{Q} + (q_2 - q_1)\mathbf{I}$. We then collect all coefficients of $\mathbf{J}, \mathbf{Q}, \mathbf{I}$ in the above product separately as

$$
\begin{aligned}
&(\eta_1\mathbf{J} + (\eta_0 - \eta_1)\mathbf{Q} + (\eta_2 - \eta_1)\mathbf{I})\,(\tilde{\eta}_1\mathbf{J} + (\tilde{\eta}_0 - \tilde{\eta}_1)\mathbf{Q} + (\tilde{\eta}_2 - \tilde{\eta}_1)\mathbf{I})\\
&= \mathbf{J}\left(\underbrace{\eta_1\tilde{\eta}_1\binom{n}{2}}_{\text{from } \mathbf{J}^2} + \underbrace{q_2(\eta_0 - \eta_1)\tilde{\eta}_1}_{\text{from } \mathbf{QJ}} + \underbrace{q_2\eta_1(\tilde{\eta}_0 - \tilde{\eta}_1)}_{\text{from } \mathbf{JQ}} + \underbrace{\eta_1(\tilde{\eta}_2 - \tilde{\eta}_1)}_{\text{from } \mathbf{JI}} + \underbrace{\tilde{\eta}_1(\eta_2 - \eta_1)}_{\text{from } \mathbf{IJ}} + \underbrace{q_1(\eta_0 - \eta_1)(\tilde{\eta}_0 - \tilde{\eta}_1)}_{\text{from } \mathbf{Q}^2}\right)\\
&+ \mathbf{Q}\left(\underbrace{(\eta_2 - \eta_1)(\tilde{\eta}_0 - \tilde{\eta}_1)}_{\text{from } \mathbf{IQ}} + \underbrace{(\eta_0 - \eta_1)(\tilde{\eta}_2 - \tilde{\eta}_1)}_{\text{from } \mathbf{QI}} + \underbrace{(q_0 - q_1)(\eta_0 - \eta_1)(\tilde{\eta}_0 - \tilde{\eta}_1)}_{\text{from } \mathbf{Q}^2}\right)\\
&+ \mathbf{I}\left(\underbrace{(\eta_2 - \eta_1)(\tilde{\eta}_2 - \tilde{\eta}_1)}_{\text{from } \mathbf{I}^2} + \underbrace{(q_2 - q_1)(\eta_0 - \eta_1)(\tilde{\eta}_0 - \tilde{\eta}_1)}_{\text{from } \mathbf{Q}^2}\right)
\end{aligned}
$$

Clearly, the coefficient of $\mathbf{J}$ and $\mathbf{Q}$ should be zero, whereas the coefficient of $\mathbf{I}$ should be equal to one to yield the identity matrix. We thus obtain the following system of equations for the coefficients of $\mathbf{J}$, $\mathbf{Q}$ and $\mathbf{I}$, respectively:

$$
\begin{aligned}
0 &= (\tilde{\eta}_0 - \tilde{\eta}_1)\cdot(q_2\eta_1 + q_1(\eta_0 - \eta_1)) && + (\tilde{\eta}_2 - \tilde{\eta}_1)\cdot\eta_1 && + \tilde{\eta}_1\cdot((\binom{n}{2} - 1 - q_2)\eta_1 + q_2\eta_0 + \eta_2)\\
0 &= (\tilde{\eta}_0 - \tilde{\eta}_1)\cdot((\eta_2 - \eta_1) + (q_0 - q_1)(\eta_0 - \eta_1)) && + (\tilde{\eta}_2 - \tilde{\eta}_1)\cdot(\eta_0 - \eta_1)\\
1 &= (\tilde{\eta}_0 - \tilde{\eta}_1)\cdot(q_2 - q_1)(\eta_0 - \eta_1) && + (\tilde{\eta}_2 - \tilde{\eta}_1)\cdot(\eta_2 - \eta_1).
\end{aligned}
$$

Solving the second equation for $\tilde{\eta}_0 - \tilde{\eta}_1$ directly yields the second condition in Lemma A.2. Inserting this result into the third equation yields

$$
\begin{aligned}
1 &= (\tilde{\eta}_2 - \tilde{\eta}_1)\cdot\left((\eta_2 - \eta_1) - \frac{(q_2 - q_1)(\eta_0 - \eta_1)^2}{\eta_2 - \eta_1 + (q_0 - q_1)(\eta_0 - \eta_1)}\right)\\
&= (\tilde{\eta}_2 - \tilde{\eta}_1)\cdot\left(\frac{(\eta_2 - \eta_1)^2 + (q_0 - q_1)(\eta_0 - \eta_1)(\eta_2 - \eta_1) - (q_2 - q_1)(\eta_0 - \eta_1)^2}{\eta_2 - \eta_1 + (q_0 - q_1)(\eta_0 - \eta_1)}\right),
\end{aligned}
$$

where solving for $\tilde{\eta}_2 - \tilde{\eta}_1$ yields the first condition in Lemma A.2. Note that the value of $\tilde{\eta}_2 - \tilde{\eta}_1$ can be explicitly computed, which implies and explicit representation of $\tilde{\eta}_0 - \tilde{\eta}_1$. Lastly, treating $\tilde{\eta}_2 - \tilde{\eta}_1$ and $\tilde{\eta}_0 - \tilde{\eta}_1$ as known and solving the first equation for $\tilde{\eta}_1$ yields the third condition in Lemma A.2. □

Having established the conditions for the inverse, we apply Lemma A.2 to $\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2 = \mu_{2,1}\mathbf{J} + (\mu_{2,0} - \mu_{2,1})\mathbf{Q} + (\mu_{2,2} - \mu_{2,1})\mathbf{I}$ to obtain $(\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1} = \tilde{\mu}_{2,1}\mathbf{J} + (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1})\mathbf{Q} + (\tilde{\mu}_{2,2} - \tilde{\mu}_{2,1})\mathbf{I}$. For the following calculations the explicit form of $\mu_{2,\ell}$ and the following relations will be used

$$
q_2 - q_1 = n - 3, \qquad q_0 - q_1 = -(n - 4), \qquad q_2 + q_1 = (n - 3)^2, \qquad \mu_{2,2} - \mu_{2,1} = \mu_{2,1}
$$

We now further simplify the terms using the structure of μ and q as

$$
\begin{aligned}
\tilde{\mu}_{2,2} - \tilde{\mu}_{2,1} &= \frac{(\mu_{2,2} - \mu_{2,1}) + (q_0 - q_1)(\mu_{2,0} - \mu_{2,1})}{(\mu_{2,2} - \mu_{2,1})^2 + (q_0 - q_1)(\mu_{2,2} - \mu_{2,1})(\mu_{2,0} - \mu_{2,1}) - (q_2 - q_1)(\mu_{2,0} - \mu_{2,1})^2},\\
&= \frac{\mu_{2,1} - (n - 4)(\mu_{2,0} - \mu_{2,1})}{\mu_{2,1}^2 - (n - 4)\mu_{2,1}(\mu_{2,0} - \mu_{2,1}) - (n - 3)(\mu_{2,0} - \mu_{2,1})^2}\\
&= \frac{(n - 3)\mu_{2,1} - (n - 4)\mu_{2,0}}{\mu_{2,1}^2 + (\mu_{2,1} - (n - 3)\mu_{2,0})(\mu_{2,0} - \mu_{2,1})}\\
&= \frac{(n - 3)\mu_{2,1} - (n - 4)\mu_{2,0}}{(n - 2)\mu_{2,1}\mu_{2,0} - (n - 3)\mu_{2,0}^2}.
\end{aligned} \tag{8}
$$

We continue with the second equation and the previous result as

$$\begin{aligned}\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1} &= -\frac{(\mu_{2,0}-\mu_{2,1})(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})}{\mu_{2,2}-\mu_{2,1}+(q_0-q_1)(\mu_{2,0}-\mu_{2,1})} \\ &= -\frac{(\mu_{2,0}-\mu_{2,1})}{\mu_{2,1}-(n-4)(\mu_{2,0}-\mu_{2,1})}\cdot(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}) \\ &= -\frac{\mu_{2,0}-\mu_{2,1}}{(n-3)\mu_{2,1}-(n-4)\mu_{2,0}}\cdot(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}) \qquad (9)\end{aligned}$$

Lastly, the third equation yields

$$\begin{aligned}\tilde{\mu}_{2,1} &= -\frac{(q_2\mu_{2,1}+q_1(\mu_{2,0}-\mu_{2,1}))(\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1})+\mu_{2,1}(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})}{(\binom{n}{2}-1-q_2)\mu_{2,1}+q_2\mu_{2,0}+\mu_{2,2}} \\ &= -\frac{((n-3)\mu_{2,1}+\binom{n-3}{2}\mu_{2,0})(\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1})+\mu_{2,1}(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})}{(\binom{n}{2}+1-\binom{n-2}{2})\mu_{2,1}+\binom{n-2}{2}\mu_{2,0}} \\ &= -\frac{(n-3)\mu_{2,1}+\binom{n-3}{2}\mu_{2,0}}{2(n-1)\mu_{2,1}+\binom{n-2}{2}\mu_{2,0}}\cdot(\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1}) - \frac{\mu_{2,1}}{2(n-1)\mu_{2,1}+\binom{n-2}{2}\mu_{2,0}}\cdot(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}). \qquad (10)\end{aligned}$$

Since our goal is to compute $\lim_{\mu_\infty\to\infty}(\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2$ we distinguish between the finite subsets, where $\mu_2(t)\neq\mu_\infty$, i.e. $2\leq|T|\leq n-2$ and the infinite subsets, where $\mu_2(t)=\mu_\infty$, i.e. $|T|=1$ or $|T|=n-1$.

T with finite weight. Clearly, for the T with $\mu_2(t)\neq\mu_\infty$, we have that $\lim_{\mu_\infty\to\infty}(\mathbf{X}_2^T\mathbf{W}_2)_{TS}=(\mathbf{X}_2^T\mathbf{W}_2)_{TS}$ and thus we can compute the limit of $(\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}$ separately. Recall that $\mu_1\propto\mu_\infty$ and $\mu_0=n-3$ does not depend on μ_∞. Hence, by Equation (8), Equation (9) and Equation (10), it follows for $\mu_\infty\to\infty$

$$\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}\to\frac{1}{n-2}, \qquad \tilde{\mu}_{2,0}-\tilde{\mu}_{2,1}\to\frac{1}{(n-3)(n-2)}, \qquad \tilde{\mu}_{2,1}\to-\frac{1}{(n-1)(n-2)}.$$

Therefore, this proves the special case of Conjecture 3.8 for $k=2$ as

$$\tilde{\mu}_{2,1}\to-\frac{1}{(n-1)(n-2)} \qquad \tilde{\mu}_{2,2}\to\frac{1}{n-1} \qquad \tilde{\mu}_{2,0}\to\frac{2}{(n-1)(n-2)(n-3)}.$$

With $\lambda(2,1)=-1/2$ it holds, $(\mathbf{J}\mathbf{X}_2^T\mathbf{W}_2)_{ST}=-\frac{1}{2}t(n-t)\mu_2(t)$, $(\mathbf{Q}\mathbf{X}_2^T\mathbf{W}_2)_{ST}=-\frac{1}{2}(t-|T\cap S|)(n-|T\cup S|)\mu_2(t)$ and $(\mathbf{I}\mathbf{X}_2^T\mathbf{W}_2)_{ST}=-\frac{1}{2}\mathbf{1}(|T\cap S|=1)\mu_2(t)$. We can thus compute

$$\begin{aligned}&\left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{ST} \\ &\overset{\mu_\infty\to\infty}{\longrightarrow}\mu_2(t)\left(\underbrace{\frac{t(n-t)}{2(n-1)(n-2)}}_{\text{from }\tilde{\mu}_{2,1}\mathbf{J}\mathbf{X}_2^T\mathbf{W}_2}-\underbrace{\frac{(t-|T\cap S|)(n-t-2+|T\cap S|)}{2(n-3)(n-2)}}_{\text{from }(\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1})\mathbf{Q}\mathbf{X}_2^T\mathbf{W}_2}-\underbrace{\frac{\mathbf{1}(|T\cap S|=1)}{2(n-2)}}_{\text{from }(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\mathbf{I}\mathbf{X}_2^T\mathbf{W}_2}\right) \\ &=\frac{(n-t-2)!(t-2)!}{(n-4)!}\cdot\begin{cases}\frac{t(n-t)}{2(n-1)(n-2)}-\frac{t(n-t-2)}{2(n-3)(n-2)}=\frac{t((n-t)(n-1)-(n-t-2)(n-3))}{2(n-1)(n-2)(n-3)}=\frac{t(t-1)}{(n-1)(n-2)(n-3)} & \text{, if } T\cap S=\emptyset \\ \frac{(n-3)t(n-t)-(n-1)(t-1)(n-t-1)-(n-1)(n-3)}{2(n-1)(n-2)(n-3)}=-\frac{(t-1)(n-t-1)}{(n-1)(n-2)(n-3)} & \text{, if } |T\cap S|=1 \\ \frac{t(n-t)}{2(n-1)(n-2)}-\frac{(t-2)(n-t)}{2(n-3)(n-2)}=\frac{(n-t)((n-3)t-(t-2)(n-1))}{2(n-1)(n-2)(n-3)}=\frac{(n-t)(n-t-1)}{(n-1)(n-2)(n-3)} & \text{, if } |T\cap S|=2\end{cases} \\ &=(-1)^{|T\cap S|}\frac{(t-|T\cap S|)!(n-t-2+|T\cap S|)!}{(n-1)!}=\frac{(-1)^{|T\cap S|}}{n-1}\binom{n-2}{t-|T\cap S|}^{-1},\end{aligned}$$

which yields the SII weights in Equation (7) and concludes the proof for this case.

T with infinite weight. Without loss of generality, we consider $T=i$, since $(\mathbf{X}_2)_{iS}=(\mathbf{X}_2)_{(N\setminus i)S}$, and $(\mathbf{X}_2)_{iS}=\lambda(2,1)=-1/2$ for all $S=ij$ with $i\neq j\in N$, and zero otherwise. It holds $(\mathbf{J}\mathbf{X}_2^T\mathbf{W}_2)_{ST}=-\frac{1}{2}(n-1)\mu_\infty$,

$(\mathbf{Q}\mathbf{X}_2^T\mathbf{W}_2)_{ST} = -\frac{1}{2}\mathbf{1}(i \notin S)(n-3)\mu_\infty$ and $(\mathbf{I}\mathbf{X}_2^T\mathbf{W}_2)_{ST} = -\frac{1}{2}\mathbf{1}(i \in S)\mu_\infty$.

$$\begin{aligned}\left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{Si} &= -\frac{1}{2}\mu_\infty\left((n-1)\tilde{\mu}_{2,1} + \mathbf{1}(i \notin S)(n-3)(\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1}) + \mathbf{1}(i\in S)(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\right)\\ &= -\frac{1}{2}\mu_\infty \cdot \begin{cases}(n-1)\tilde{\mu}_{2,1} + (n-3)(\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1}) & \text{, if } i \notin S\\ (n-1)\tilde{\mu}_{2,1} + (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}) & \text{, if } i \in S\end{cases}\end{aligned}$$

We now first explicitly compute $\tilde{\mu}_{2,1}$ with Equation (10) and Equation (9) as

$$\begin{aligned}\tilde{\mu}_{2,1} &\overset{(10)}{=} -\frac{(n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} \cdot (\tilde{\mu}_{2,0}-\tilde{\mu}_{2,1}) - \frac{\mu_{2,1}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} \cdot (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\\ &\overset{(9)}{=} \left(\frac{(n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} \cdot \frac{\mu_{2,0}-\mu_{2,1}}{(n-3)\mu_{2,1}-(n-4)\mu_{2,0}} - \frac{\mu_{2,1}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}}\right) \cdot (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\\ &= \left(\frac{((n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}) \cdot (\mu_{2,0}-\mu_{2,1}) - \mu_{2,1}((n-3)\mu_{2,1}-(n-4)\mu_{2,0})}{(2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}) \cdot ((n-3)\mu_{2,1}-(n-4)\mu_{2,0})}\right) \cdot (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\\ &= \left(\frac{((n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}) \cdot (\mu_{2,0}-\mu_{2,1}) - \mu_{2,1}((n-3)\mu_{2,1}-(n-4)\mu_{2,0})}{(2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}) \cdot ((n-3)\mu_{2,1}-(n-4)\mu_{2,0})}\right) \cdot (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\\ &= \left(\frac{-2(n-3)\mu_{2,1}^2 + ((n-4)+(n-3)-\binom{n-3}{2})\mu_{2,1}\mu_{2,0} + \binom{n-3}{2}\mu_{2,0}^2}{(2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}) \cdot ((n-3)\mu_{2,1}-(n-4)\mu_{2,0})}\right) \cdot (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\\ &=: \frac{\gamma^\uparrow}{\gamma^\downarrow} \cdot (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}) \qquad (11)\end{aligned}$$

Utilizing the new notation, we obtain for $i \in S$

$$\begin{aligned}\left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{Si} &= -\frac{1}{2}\mu_\infty((n-1)\tilde{\mu}_{2,1} + (\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})) \overset{(11)}{=} -\frac{1}{2}\mu_\infty((n-1)\frac{\gamma^\uparrow}{\gamma^\downarrow}+1)(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\\ &= -\frac{1}{2}\mu_\infty\frac{(n-1)\gamma^\uparrow + \gamma^\downarrow}{\gamma^\downarrow}(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}).\end{aligned}$$

In the following, we use $\approx$ to indicate that terms grow similarly for $\mu_\infty \to \infty$. We further observe that the terms with $\mu_{2,1}^2$ vanish in $(n-1)\gamma^\uparrow + \gamma^\downarrow$, and

$$\begin{aligned}(n-1)\gamma^\uparrow + \gamma^\downarrow &\approx \mu_{2,1}\left((n-1)((n-4)+(n-3)-\binom{n-3}{2})\mu_{2,0} + \mu_{2,0}(\binom{n-2}{2}(n-3) - 2(n-1)(n-4))\right)\\ &= \mu_{2,1}\mu_{2,0}((n-1)-(n-1)\binom{n-3}{2} + (n-3)\binom{n-2}{2})\\ &= \mu_{2,1}\mu_{2,0}((n-1)-(n-3)(\binom{n-3}{2}-\binom{n-2}{2}) - 2\binom{n-3}{2})\\ &= \mu_{2,1}\mu_{2,0}((n-1)-(n-3)(n-3)-(n-3)(n-2))\\ &= 2\mu_{2,1}\mu_{2,0}(n-2).\end{aligned}$$

Clearly, $\gamma^\downarrow \approx 2(n-1)(n-3)\mu_{2,1}^2$ are the dominating terms. Hence, with $\mu_{2,1} \approx \frac{1}{2}\mu_\infty$, we have $\frac{1}{2}\mu_\infty\frac{\mu_{2,1}}{\mu_{2,1}^2} \to 1$ and with $\mu_{2,0} = (n-3)$ we compare the coefficients of $\mu_{2,1}$ to obtain

$$\lim_{\mu_\infty \to \infty}\left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{Si} = -\underbrace{\frac{2(n-3)(n-2)}{2(n-1)(n-3)}}_{\text{from } \frac{1}{2}\mu_\infty\frac{(n-1)\gamma^\uparrow+\gamma^\downarrow}{\gamma^\downarrow}} \cdot \underbrace{\frac{1}{n-2}}_{\text{from } \tilde{\mu}_{2,2}-\tilde{\mu}_{2,1}} = -\frac{1}{n-1} = \frac{(-1)^{|T\cap S|}}{n-1}\binom{n-2}{t-|T\cap S|}^{-1},$$

which proves the convergence to the SII weight for the case where $T = i$ and $i \in S$.

To prove the case for $T = i$ and $i \notin S$, we first represent $\tilde{\mu}_{2,1}$ with $\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}$. Using the reverse relation in Equation (9) and Equation (10), we obtain

$$\begin{aligned}
\tilde{\mu}_{2,1} &\overset{(10)}{=} -\frac{(n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} \cdot (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}) - \frac{\mu_{2,1}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} \cdot (\tilde{\mu}_{2,2} - \tilde{\mu}_{2,1}) \\
&\overset{(9)}{=} \left(-\frac{(n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} + \frac{\mu_{2,1}}{2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}} \cdot \frac{(n-3)\mu_{2,1} - (n-4)\mu_{2,0}}{\mu_{2,0} - \mu_{2,1}} \cdot \right) (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}) \\
&= \frac{-(\mu_{2,0} - \mu_{2,1}) \cdot ((n-3)\mu_{2,1} + \binom{n-3}{2}\mu_{2,0}) + \mu_{2,1} \cdot ((n-3)\mu_{2,1} - (n-4)\mu_{2,0})}{(2(n-1)\mu_{2,1} + \binom{n-2}{2}\mu_{2,0}) \cdot (\mu_{2,0} - \mu_{2,1})} \cdot (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}) \\
&= \frac{\mu_{2,1}^2 2(n-3) - \mu_{2,1}\mu_{2,0}((n-3) - \binom{n-3}{2} + (n-4)) - \mu_{2,0}^2\binom{n-3}{2}}{-\mu_{2,1}^2 2(n-1) + \mu_{2,1}\mu_{2,0}(2(n-1) - \binom{n-2}{2}) + \mu_{2,0}^2\binom{n-2}{2}} \cdot (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}) \\
&= \frac{\mu_{2,1}^2 2(n-3) - \mu_{2,1}\mu_{2,0}((n-3) - \binom{n-3}{2} + (n-4)) - \mu_{2,0}^2\binom{n-3}{2}}{-\mu_{2,1}^2 2(n-1) + \mu_{2,1}\mu_{2,0}(2(n-1) - \binom{n-2}{2}) + \mu_{2,0}^2\binom{n-2}{2}} \cdot (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}) \\
&=: \frac{\tau^{\uparrow}}{\tau^{\downarrow}} \cdot (\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}). \qquad (12)
\end{aligned}$$

Again, utilizing the new notation, we obtain for $i \notin S$

$$\begin{aligned}
\left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{Si} &= -\frac{1}{2}\mu_\infty((n-1)\tilde{\mu}_{2,1} + (n-3)(\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1})) \\
&\overset{(12)}{=} -\frac{1}{2}\mu_\infty((n-1)\frac{\tau^{\uparrow}}{\tau^{\downarrow}} + n-3)(\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}) \\
&= -\frac{1}{2}\mu_\infty\frac{(n-1)\tau^{\uparrow} + (n-3)\tau^{\downarrow}}{\tau^{\downarrow}}(\tilde{\mu}_{2,0} - \tilde{\mu}_{2,1}).
\end{aligned}$$

We observe again that the terms of $\mu_{1,2}^2$ cancel in $(n-1)\tau^{\uparrow} + (n-3)\tau^{\downarrow}$ and thus the coefficients of $\mu_{2,1}$ are the dominating terms for $\mu_\infty \to \infty$. Therefore,

$$\begin{aligned}
(n-1)\tau^{\uparrow} + (n-3)\tau^{\downarrow} &\approx \mu_{2,1}\mu_{2,0}(-(n-1)((n-3) - \binom{n-3}{2} + (n-4)) + (n-3)(2(n-1) - \binom{n-2}{2})) \\
&= \mu_{2,1}\mu_{2,0}(\underbrace{(n-1)(n-3) - (n-1)(n-4)}_{=n-1} + \underbrace{\binom{n-3}{2}((n-1) - (n-3)) - (n-3)(n-3))}_{=(n-3)(n-4)-(n-3)(n-3)=-(n-3)} \\
&= 2\mu_{2,1}\mu_{2,0}.
\end{aligned}$$

Clearly the denominating terms in $\tau^{\downarrow}$ are $\tau^{\downarrow} \approx -\mu_{2,1}^2 2(n-1)$. Hence, with $\mu_{2,1} \approx \frac{1}{2}\mu_\infty$, we have $\frac{1}{2}\mu_\infty\frac{\mu_{2,1}}{\mu_{2,1}^2} \to 1$ and with $\mu_{2,0} = (n-3)$ we compare the coefficients of $\mu_{2,1}$ to obtain

$$\lim_{\mu_\infty \to \infty} \left((\mathbf{X}_2^T\mathbf{W}_2\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}_2\right)_{Si} = -\frac{2(n-3)}{-2(n-1)} \cdot \underbrace{\frac{1}{(n-2)(n-3)}}_{\text{from } \tilde{\mu}_{2,0}-\tilde{\mu}_{2,1}} = \frac{1}{(n-1)(n-2)} = \frac{(-1)^{|T\cap S|}}{n-1}\binom{n-2}{t-|T\cap S|}^{-1},$$

which proves the convergence to the SII weight for $T = i$ and $i \notin S$, and finishes the proof. □

B. Algorithms and Ground Truth Calculations

In this section, we give further details on the implemented algorithms, the default parameters of KernelSHAP-IQ and the computation of GT SII values for the synthetic SOUM.

B.1. KernelSHAP-IQ Default Parameters

We use the following default configurations in our experiments:

- **Sampling weights:** We use $p^*(T) \propto \mu_1(t)$, similar to KernelSHAP (Covert & Lee, 2021), i.e.

  $$q(t) := p^*(|T| = t) \propto \binom{n}{t}\mu_1(t) \propto \frac{1}{t(n-t)}.$$

  We further use similar sampling weights for the baseline methods, if applicable, i.e. SHAP-IQ (Fumagalli et al., 2023) and SVARM-IQ (Kolpaczki et al., 2024b).

- **Border-Trick:** We use the *border-trick* (Fumagalli et al., 2023; Lundberg & Lee, 2017) to split the sampling procedure in a deterministic and a sampling part, cf. Appendix B.3. We also apply this method to the baseline algorithms, were it is applicable, i.e. SHAP-IQ (Fumagalli et al., 2023) and SVARM-IQ (Kolpaczki et al., 2024b).

- **Infinite Weight:** We set $\mu_\infty = 10^6$, in line with KernelSHAP (Lundberg & Lee, 2017).

B.2. Inconsistent KernelSHAP-IQ

The inconsistent KernelSHAP-IQ procedure is similar to KernelSHAP-IQ. However, there is no iterative computation involved and the weights do not change. We outline the pseudocode in Algorithm 2.

Algorithm 2 Inconsistent KernelSHAP-IQ

```
Require: order $k$, sampling weights $q$, budget $b$.
$\{T_i\}_{i=1,\dots,b}, \{w_T\}_{T=T_1,\dots,T_b} \leftarrow \textsc{Sample}(q, b)$
$\mathbf{y}_1 \leftarrow [\nu(T_1), \dots, \nu(T_b)]^T$
for $T \in \{T_i\}$ and $1 \leq |S| \leq k$ do
    $(\hat{\mathbf{X}}_{\leq k})_{TS} \leftarrow \lambda(|S|, |T \cap S|)$        ▷ Bernoulli weighting
    $(\hat{\mathbf{W}}^*_{\leq k})_{TT} \leftarrow \mu_1(t) \cdot w_T$        ▷ weight adjustment
end for
$\hat{\phi}_{\leq k} \leftarrow \textsc{SolveWLS}(\hat{\mathbf{X}}_{\leq k}, \hat{\mathbf{y}}_{\leq k}, \hat{\mathbf{W}}^*_{\leq k})$
$\hat{\phi}_1, \dots, \hat{\phi}_k \leftarrow \textsc{Unstack}(\hat{\phi}_{\leq k})$
$\hat{\Phi}_k \leftarrow \textsc{AggregateSII}(\hat{\phi}_1, \dots, \hat{\phi}_k)$        ▷ compute $k$-SII
return $k$-SII estimates $\hat{\Phi}_k$, SII estimates $\hat{\phi}_{\leq k}$
```

B.3. Sampling Algorithm

In this section, we describe our sampling approach. We make use of the *border-trick* (Fumagalli et al., 2023), which computes the low- and high-cardinality subsets explicitly without sampling, if the expected number of subsets is higher than the number of subsets of that size. The sampling procedure is split in a *deterministic* ($t < q_0$ and $t > n - q_0$) and a *sampling* part ($q_0 \leq t \leq n - q_0$). The method takes a (symmetric) sampling weight vector $q \geq 0$ and budget $b > 0$ and returns b distinct subsets with *subset weights* w_T for each subset adjusted for the sampling distribution to readily apply the KernelSHAP-IQ weights at a later step.

Consider first the initial probability distribution $p^*(T) \propto q(t)$. It is clear that this distribution is explicitly given as

$$p^*(T) := \frac{q(t)}{\binom{n}{t} \cdot \sum_{\ell=0}^{n} q(\ell)}, \text{ where } p^*(|T| = t) = \frac{q(t)}{\sum_{\ell=0}^{n-0} q(\ell)}.$$

The expected number of subsets of a given size t is thereby given as $p(|T| = t) \cdot b$, where we consider the subset size t in the deterministic part, if the expected number of subsets exceeds the total number of subsets of that size $\binom{n}{t}$. The procedure that splits the subset sizes in the deterministic and sampling part based on this notion is outlined in Algorithm 4, where q_0 is iteratively increased and finally returned. Given the sampling order q_0, we can re-define the sampling probabilities as

$$p^*_{q_0}(T) := \frac{q(t)}{\binom{n}{t} \cdot \sum_{\ell=q_0}^{n-q_0} q(\ell)}, \text{ where } p^*_{q_0}(|T| = t) = \frac{q(t)}{\sum_{\ell=q_0}^{n-q_0} q(\ell)},$$

where $p^*_{q_0}$ is now a probability distribution over $\mathcal{T}_{q_0}$, i.e. subsets of size $q_0 \leq t \leq n - q_0$. Recall that our goal is to approximate a sum over squared losses $\tilde{\nu}(T)$ weighted by the KernelSHAP-IQ weights $\mu(t)$. We can re-write this sum as

$$\sum_{T\subseteq N} \mu(t)\tilde{\nu}(T) = \sum_{T\subseteq N}^{T\notin\mathcal{T}_{q_0}} \mu(t)\tilde{\nu}(T) + \mathbb{E}_{T\sim p^*_{q_0}}\left[\frac{\mu(t)\tilde{\nu}(T)}{p^*_{q_0}(T)}\right] \approx \sum_{T\subseteq N}^{T\notin\mathcal{T}_{q_0}} \mu(t)\tilde{\nu}(T) + \frac{1}{n_{\text{samples}}} \sum_{\ell=1}^{n_{\text{samples}}} \frac{\mu(t_\ell)\tilde{\nu}(T_\ell)}{p^*_{q_0}(T_\ell)},$$

where the n_{samples} Monte Carlo samples $T_1, \ldots, T_{n_{\text{samples}}}$ are drawn according to $p^*_{q_0}$. In practice, note it is easy to sample from this distribution by sampling a subset size according to $p^*_{q_0}(|T| = t)$ and then sample a subset T of that size uniformly with probability $\binom{n}{t}^{-1}$. Clearly, the sampling procedure should therefore return the sampling probabilities

Deterministic part: $w_T \equiv 1$ **Sampling part:** $w_T = \dfrac{1}{n_{\text{samples}} \cdot p^*_{q_0}(T)}$.

If a subset is sampled multiple times, then clearly the budget should be decreased, as the game only needs to be evaluated once. As a consequence, we simply increase the weight proportionally and leave the budget unchanged. The full algorithm is outlined in Algorithm 3.

Algorithm 3 Sampling

```
Require: Budget $b > 0$, sampling weights $\{q(t)\}_{t=0,\ldots,n}$
$q_0 \leftarrow$ SamplingOrder$(b, q)$
$i \leftarrow 1$
for $T \subseteq N$ with $T \notin \mathcal{T}_{q_0}$ do                          ▷ deterministic part
    $T_i \leftarrow T$
    $w_T \leftarrow 1$                                                  ▷ no weighting adjustment
    $b \leftarrow b - 1$                                                ▷ reduce budget
end for
$q_{\text{sampling}}(t) \leftarrow q(t) / \sum_{\ell=q_0}^{n-q_0} q(\ell)$ for $q_0 \leq t \leq n - q_0$    ▷ sampling weight normalization, i.e. $p^*_{q_0}(|T| = t)$
$n_{\text{samples}} \leftarrow b$                                          ▷ store total number of subsets sampled
while $b > 0$ do                                                       ▷ sampling part
    $t \leftarrow$ Sample according to $q_{\text{sampling}}$
    $T \leftarrow$ Sample uniformly of size $t$
    $p(T) \leftarrow q_{\text{sampling}}(t) \cdot \binom{n}{t}^{-1}$                    ▷ probability to draw this subset
    if $T \notin \{T_i\}$ then                                           ▷ Subset is new
        $w_T \leftarrow (n_{\text{samples}} \cdot p(T))^{-1}$
        $b \leftarrow b - 1$                                            ▷ reduce budget
    else                                                               ▷ Subset is not new
        $w_T \leftarrow w_T + (n_{\text{samples}} \cdot p(T))^{-1}$            ▷ increase weighting for each occurrence of $T$
    end if
end while
return $\{T_i\}_{i=1,\ldots,b}$, $\{w_T\}_{T=T_1,\ldots,T_b}$
```

B.4. Solving WLS

The WLS problem can be solved analytically as described in Algorithm 5.

Algorithm 4 Compute Sampling Order

```
Require: Budget $b > 0$, sampling weights $\{q(t)\}_{t=0,\dots,n}$
initialize $q_0 = 0$
for $t = 0, \dots, \textsc{Floor}(n/2)$ do
    $q(t) \leftarrow q(t) / \sum_{\ell=q_0}^{n-q_0} q(\ell)$                         ▷ weight normalization
    if $b \cdot q(t) \geq \binom{n}{t}$ and $b \cdot q(n-t) \geq \binom{n}{t}$ then   ▷ compare expected number of subsets with total number of subsets
        $q_0 \leftarrow q_0 + 1$                                                     ▷ increase $q_0$
        $b \leftarrow b - 2\binom{n}{t}$
    else
        break
    end if
end for
return $q_0$
```

Algorithm 5 SolveWLS

```
Require: Data $\mathbf{X}$, weights $\mathbf{W}$, response $\mathbf{y}$
$\mathbf{A} \leftarrow (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$                 ▷ precision matrix
$\phi \leftarrow \mathbf{A}\mathbf{X}^T\mathbf{W}\mathbf{y}$
return $\phi$
```

B.5. Subset Weights SII

We briefly describe the weighting for SII. According to Fumagalli et al. (2023), the SII is represented as

$$\phi^{\text{SII}}(S) = \sum_{T \subseteq N} (-1)^{s-|T \cap S|} m(t - |T \cap S|), \text{ where } m(t) := \frac{(n-t-s)!t!}{(n-s+1)!}.$$

The weights are therefore assigned to every combinations T, S as outlined in Algorithm 6.

Algorithm 6 SIIWeight

```
Require: Subset $T \subseteq N$, interaction $S \subseteq N$
$t, s, r \leftarrow |T|, |S|, |T \cap S|$
$\omega \leftarrow (-1)^{s-r} \frac{(n-t-s+r)!(t-r)!}{(n-s+1)!}$
return $\omega$
```

B.6. Aggregate SII to k-SII

This section describes the k-SII weights, i.e. the aggregation of SII to k-SII. It has been shown in Appendix A.2 in (Bordt & von Luxburg, 2023) that k-SII can be explicitly represented as

$$\Phi_k(S) = \sum_{S \subseteq \tilde{S} \subseteq N}^{|\tilde{S}| \leq k} B_{|\tilde{S}|-|S|} \phi^{\text{SII}}(\tilde{S}) = \sum_{\tilde{S} \subseteq N}^{|\tilde{S}| \leq k} \mathbf{1}_{S \subseteq \tilde{S}} B_{|\tilde{S}|-|S|} \phi^{\text{SII}}(\tilde{S}) \text{ for } 1 \leq |S| \leq k.$$

Hence, we can aggregate SII to k-SII as outlined in Algorithm 7.

Algorithm 7 AggregateSII

```
Require: SII estimates $\hat{\phi}_1, \dots, \hat{\phi}_k$
$(\mathbf{Z}_k)_{S\tilde{S}} \leftarrow \mathbf{1}_{S \subseteq \tilde{S}} B_{|\tilde{S}|-|S|}$ for $1 \leq |S|, |\tilde{S}| \leq k$
$\hat{\Phi}_k \leftarrow \mathbf{Z}_k \cdot [\hat{\phi}_1, \dots, \hat{\phi}_k]^T$
return $\hat{\Phi}_k$
```

B.7. Baseline Methods: SHAP-IQ, Permutation Sampling and SVARM-IQ

The pseudocode and implementations can be found in the corresponding paper. Permutation sampling (Tsai et al., 2023; Fumagalli et al., 2023) is described for SII as an extension to permutation sampling of the SV (Castro et al., 2009). SHAP-IQ (Fumagalli et al., 2023) is implemented using similar sampling weights as KernelSHAP-IQ and extends Unbiased KernelSHAP (Covert & Lee, 2021) to interactions. SVARM-IQ (Kolpaczki et al., 2024b) extends SVARM (Kolpaczki et al., 2024a) to interactions.

B.8. Analytic Solution for SII of SOUMs

Proposition B.1. *For a unanimity game $\nu_R(T) := \mathbf{1}_{R\subseteq T}$ with $R \subseteq N$, the SII for any $S \subseteq N$ can be explicitly computed as*

$$\phi^{\textit{SII}}_{\nu_R}(S) = \frac{\mathbf{1}_{S\subseteq R}}{r-s+1}.$$

Hence for a SOUM, defined as $\nu^{\textit{SOUM}}(T) := \sum_{m=1}^{M} a_m \nu_{R_m}(T)$ with subsets $R_1,\dots,R_M \subseteq N$ and coefficients $a_1,\dots,a_M \in \mathbb{R}$, it follows

$$\phi^{\textit{SII}}_{\nu^{\textit{SOUM}}}(S) = \sum_{m=1}^{M} a_m \frac{\mathbf{1}_{S\subseteq R_m}}{r_m - s + 1}.$$

Proof. Note that due to linearity the second result follows immediately from the SII of the unanimity game. We thus aim to compute the SII for a unanimity game ν_R with subset $R \subseteq N$. In the following, we give two different proofs.

Proof via Möbius transform and conversion formula. It is well-known in cooperative game theory, cf. p.54 in (Grabisch, 2016), that the Möbius transform of a unanimity game is

$$a_{\nu_R}(S) = \sum_{T\subseteq S} (-1)^{s-t} \nu_R(T) = \mathbf{1}_{R=S}.$$

Furthermore, the result follows immediately from the conversion formula, cf. Table 3 in (Grabisch et al., 2000), as

$$\phi^{\text{SII}}_{\nu_R}(S) = \sum_{T\supseteq S} \frac{1}{t-s+1} a_{\nu_R}(T) = \sum_{T\subseteq N} \frac{\mathbf{1}_{S\subseteq T}}{t-s+1} a_{\nu_R}(T) = \mathbf{1}_{R=T} \frac{\mathbf{1}_{S\subseteq T}}{t-s+1} = \frac{\mathbf{1}_{S\subseteq R}}{r-s+1}.$$

Proof via computation. We can also give an alternative analytical proof. We compute the discrete derivatives of ν_R for $S \subseteq N$ and $T \subseteq N \setminus S$ as

$$\begin{aligned}\Delta_S(T) &= \sum_{L\subseteq S} (-1)^{s-\ell} \nu_R(T\cup L) = \sum_{L\subseteq S} (-1)^{s-\ell} \mathbf{1}_{R\subseteq T\cup L} = \mathbf{1}_{R\setminus S\subseteq T} \sum_{L\subseteq S} (-1)^{s-\ell} \mathbf{1}_{R\cap S\subseteq L} \\ &= \mathbf{1}_{R\setminus S\subseteq T} \sum_{\ell=|R\cap S|}^{s} (-1)^{s-\ell} \binom{s-|R\cap S|}{\ell-|R\cap S|} = \mathbf{1}_{R\setminus S\subseteq T} \underbrace{\sum_{\ell=0}^{s-|R\cap S|} (-1)^{s-\ell-|R\cap S|} \binom{s-|R\cap S|}{\ell}}_{=0,\text{ except for } |R\cap S|=s} = \mathbf{1}_{R\setminus S\subseteq T} \mathbf{1}_{S\subseteq R}.\end{aligned}$$

Hence, with $q := |R \setminus S|$, the SII is computed as

$$\begin{aligned}
\phi^{\text{SII}}(S) &= \sum_{T \subseteq N \setminus S} \frac{(n-t-s)!t!}{(n-s+1)!} \Delta_S(T) && = \mathbf{1}_{S \subseteq R} \sum_{T \subseteq N \setminus S} \frac{(n-t-s)!t!}{(n-s+1)!} \mathbf{1}_{R \setminus S \subseteq T} \\
&= \mathbf{1}_{S \subseteq R} \sum_{t=q}^{n-s} \frac{(n-t-s)!t!}{(n-s+1)!} \binom{n-s-q}{t-q} && = \mathbf{1}_{S \subseteq R} \sum_{t=0}^{n-s-q} \frac{(t+q)!}{(n-s+1)!} \frac{(n-s-q)!}{t!} \\
&= \mathbf{1}_{S \subseteq R} \frac{(n-s-q)!}{(n-s+1)!} \sum_{t=0}^{n-s-q} \frac{(t+q)!}{q!} && = \mathbf{1}_{S \subseteq R} \frac{(n-s-q)!}{(n-s+1)!} q! \sum_{t=0}^{n-s-q} \binom{t+q}{q} \\
&= \mathbf{1}_{S \subseteq R} \frac{(n-s-q)!}{(n-s+1)!} q! \sum_{t=0}^{n-s-q} \binom{t+q}{q} && = \mathbf{1}_{S \subseteq R} \frac{(n-s-q)!}{(n-s+1)!} q! \binom{n-s+1}{n-s-q} \\
&= \mathbf{1}_{S \subseteq R} \frac{1}{q+1} && = \frac{\mathbf{1}_{S \subseteq R}}{r-s+1},
\end{aligned}$$

where we used the hockey-stick identity in the second last row to compute the sum and $q = r - s$, if $S \subseteq R$. □

B.9. Intuition about the KernelSHAP-IQ Weights

The intuition behind $\mathbf{W}_k$ is based on common factors in two representations of SII: One being the solution of the WLS problem from Equation (3)

$$\phi_k^{\text{SII}} = (\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{W}_k \cdot \mathbf{y}_k,$$

and the other being the representation of Conjecture 3.9 and by Fumagalli et al. (2023)

$$\phi_k^{\text{SII}} = \mathbf{Q}_k \cdot \mathbf{y}_k.$$

In detail, if we consider the final weight of a single subset $T \subseteq N$ with $k \leq |T| \leq n-k$ and all interactions $S \subseteq N$ of order $|S| = k$, then the common factors for all S are according to the S-row and T-column of $\mathbf{Q}_k$, i.e. $(-1)^{k-|T \cap S|} m_k(t - |T \cap S|)$, determined by the common factors in $m_k(t-0), \ldots, m_k(t-k)$. By definition of $m_k(t) = \frac{(n-k-t)!t!}{(n-k+1)!}$, we have $m_k(t-\ell) = \frac{(n-k-t+\ell)!(t-\ell)!}{(n-k+1)!}$ and thus the common factors in $\mathbf{Q}_k$ for all S-rows given a T-column are $\frac{(n-k-t)!(t-k)!}{(n-k+1)!} \propto \binom{n-2k}{t-k}^{-1}$. On the other hand, by Equation (3), the matrix $(\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1}$ is independent of the subset T, and hence $\mu_k(t)$, as it contains one row and one column for each interaction S of order k. This implies

$$((\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{W}_k)_{ST} \propto \mu_k(t)$$

for all interactions S of order k. Combining both proportionality results, we conclude that $\mu_k(t) \propto \binom{n-2k}{t-k}^{-1}$ is a suitable candidate. Note that this reasoning works for subsets with $k \leq |T| \leq n-k$.

C. Experimental Setup and Reproducibility

This section contains additional information about the setup and reproducibility of our empirical evaluation. Appendix C.1 contains additional information about the models, datasets and explanation tasks used for evaluation as summarized in Table 2. Appendix C.2 contains additional information about the environmental impact and computational effort of the empirical evaluation. All experiments can be reproduced via the technical supplement available at https://github.com/FFmgll/KernelSHAP_IQ_Supplementary_Materialwhich will be made fully publicly available.

C.1. Model, Datasets and Task Descriptions

This section contains detailed information about the models and datasets used for benchmarking the approximation quality of KernelSHAP-IQ and the available baseline algorithms. To increase the comparability, we conduct our empirical evaluation of the approximation quality on settings and tasks presented in the literature (Kolpaczki et al., 2024b; Fumagalli et al., 2023; Muschalik et al., 2024; Sundararajan et al., 2020; Tsai et al., 2023). For further questions regarding the technical details, we further refer to the technical supplement at https://github.com/FFmgll/KernelSHAP_IQ_Supplementary_Materialwhich will be made fully publicly available.

C.1.1. Sum Of Unanimity Models (SOUM)

The sum of unanimity models (SOUMs) are used in different empirical evaluations of Shapley-based interaction estimators (Fumagalli et al., 2023; Kolpaczki et al., 2024b; Tsai et al., 2023). This synthetic model class can be used to create explanation tasks with differing complexity while allowing to compute GT interaction scores analytically even for higher player counts. For a given player set $N := \{1, \dots, n\}$ with n many players, we sample $M = 50$ interactions $R_1, \dots, R_M \subseteq N$ from $\mathcal{P}(N)$. Next, we draw for each interaction subset R_m a coefficient $a_m \in [0, 1]$ uniformly at random. The value function is then defined as $\nu(T) = \sum_{m=1}^{M} a_m \cdot \nu_{R_m}(T)$ for all coalitions $T \subseteq N$ with the unanimity game $\nu_{R_m}(T) := \mathbf{1}_{T \supseteq R_m}$ Similar to Tsai et al. (2023), we limit the highest interaction size $|R_m| \leq 4$. Further, we randomly set two features to be non-informative (dummy player) which are never part of any interaction subset.

C.1.2. Language Model (LM)

Language models (LMs) are widely investigated with Shapley interactions (Fumagalli et al., 2023; Kolpaczki et al., 2024b; Tsai et al., 2023; Sundararajan et al., 2020) highlighting that explanations based on interactions are more expressive than word-level explanations. Similar to the related work, we also investigate KernelSHAP-IQ's approximation quality in a LM scenario. The LM used in our experiments is a pre-trained DistilBert (Sanh et al., 2019) sentiment analysis model. The LM used is available at https://huggingface.co/lvwerra/distilbert-imdb via the transformers API (Wolf et al., 2020). The transformer was pre-trained on the IMDB movie review dataset (Maas et al., 2011; Lhoest et al., 2021). Provided a tokenized text input, the model predicts the sentiment of the input with a sentiment score ranging from negative -1 to positive 1. For explanation purposes, we remove features (tokens) from the tokenized representation of the input.

C.1.3. Convolutional Neural Network (CNN)

Similar to Fumagalli et al. (2023); Kolpaczki et al. (2024b), we evaluate the approximation quality of KernelSHAP-IQ and the available baselines on a convolutional neural network (CNN) image-input explanation task. The CNN is a ResNet18 (He et al., 2016) fitted on the ImageNet (Deng et al., 2009) dataset. The CNN is available at https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html via pytorch (Paszke et al., 2019). The task is to explain the predicted label's probability for random ImageNet images. Individual pixels are grouped together as superpixels with SLIC (Achanta et al., 2012). Missing features (superpixels) are set to the gray as a method of mean-imputation.

C.1.4. Vision Transformer (ViT)

The vision transformer (ViT) setting presented in Kolpaczki et al. (2024b) is similar to the CNN as the task is to explain random ImageNet images in terms of the predicted class probability. The specific ViT model operates on "words" of 32x32 pixels image patches. The ViT model is available at https://huggingface.co/google/vit-base-patch32-384 via the transformers API (Wolf et al., 2020). To create a player set of 16 players (computation of GT SII values already requires $2^{16} = 65\,536$ coalitions) the smaller 32x32 image patches are grouped together into 96x96 pixels super-patches (3x3 original sized patches make up one larger patch). Similar to the LM, absent players (patches) are removed on the tokenized representation of the input. The value of a coalition is the ViT's predicted class probability for the class which has the highest probability provided the grand coalition (the unmodified image with no patches removed).

C.1.5. Bike Rental (BR)

Based on Muschalik et al. (2024), the bike rental (BR) setting is based on the *bike* regression dataset (Fanaee-T & Gama, 2014), where the goal is to predict the amount of rented bikes based on features like weather conditions or time of day. The dataset is retrieved from `openml` (Feurer et al., 2020) with *42712* as the dataset identifier. We encode categorical features with ordinal values and scale numeric features. We further logarithmize the target variable to base ten. Based on this dataset, we train an XGBoost model (Chen & Guestrin, 2016) and explain randomly sampled local instances by removing numerical and categorical features through mean and mode imputation, respectively.

C.1.6. California Housing (CH)

Like Muschalik et al. (2024), we retrieve the California housing (CH) dataset (Kelley Pace & Barry, 1997) through `scikit-learn` library (Pedregosa et al., 2011). The CH regression dataset contains information about property prices (the continuous target variable) and corresponding property attributes including the location in terms of latitude and longitude. We standardize all features and, likewise to the BR dataset, logarithmize the target variable to base ten. We train a small NN with `pytorch` on the regression task and explain randomly sampled local instances the model by removing missing features through mean imputation. For the illustration in Figure 2, we fit a gradient boosted tree with `scikit-learn`.

C.1.7. Adult Census (AC)

The adult census (AC) classification dataset (Kohavi, 1996) contains socio-demographic features of individuals paired with their income levels. Similar to Muschalik et al. (2024), we retrieve the AC dataset via `openml` and *1590* as the dataset identifier. We transform the dataset by imputing missing numerical attributes with median values and then apply standard scaling. Categorical features are encoded with ordinal values. We train a random forest (RF) classifier from `scikit-learn` on this dataset and explain randomly sampled local instances by removing missing features through mean and mode imputation for numerical and categorical features, respectively.

C.2. Computational Effort

The computational effort required for evaluating KernelSHAP-IQ in comparison to all available baselines is modest. The main computational burden lies in the pre-computation of GT values and queries to underlying black box models. The largest effort for pre-computation of GT values stems from the ViT with $d = 16$ players requiring $2^{16} = 65\,536$ queries to the ViT model for each input image. To alleviate this computational burden, we pre-compute the worth of each coalition for *each* instance of *each* benchmark task, except the synthetic SOUMs. Hence, for each run of the different estimators the pre-computed coalition values can be looked up from a file. The code for recreating or extending these look-up files is part of the technical supplement at `https://github.com/FFmgll/KernelSHAP_IQ_Supplementary_Material`. All benchmarks are performed on a single a Dell XPS 15 9510 Laptop with an Intel i7-11800H clocking at 2.30GHz. The experiments consumed around 500 CPU hours.

D. Additional Empirical Results

D.1. Runtime Analysis

Next to the experiments regarding approximation quality, we also conducted a small runtime analysis of KernelSHAP-IQ and all available baseline algorithms. The results are summarized in Figure 5. We run KernelSHAP-IQ, SHAP-IQ, SVARM-IQ, and permutation sampling on the LM and an example sentence consisting of 14 words. We let the approximation algorithms estimate the SII values of order $l = 2$. The main computational cost results from the model evaluations (accesses to ν), which is bounded by the model's inference time. From Figure 5 it is apparent that with increasing the amount of allowed model accesses, the runtime scales linearly for all approximation algorithms and all estimators perform equally. Further, we evaluate the runtime of the different approximation methods jointly with their performance. Table 3 shows the mean (averaged over 10 runs) runtime of each approximator to reach certain MSE levels. For *CH*, KernelSHAP-IQ and Inconistent KernelSHAP-IQ are most efficient in terms of model calls to reach the desired MSE error levels. The runtime is similar for the other baselines. While Inconsistent KernelSHAP-IQ fails to reach the desired error levels for *LM*, KernelSHAP-IQ performs equally well with SVARM-IQ for *LM*. Permutation sampling generally has a fast runtime, but does not achieve good estimation qualities in sensible time.

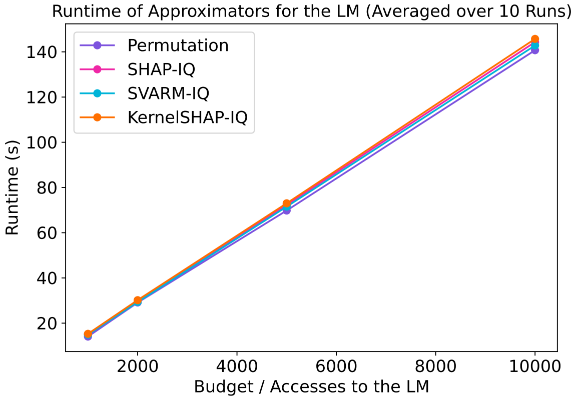

Figure 5. Runtime analysis of KernelSHAP-IQ and baseline algorithms for calculating $l = 2$ SII scores for an example sentence with $n = 14$ words and the LM. For each approximator we evaluate 10 independent runs. The shaded bands corresponds to the SEM.

Table 3. Mean runtime of each approximator to reach certain MSE error levels in terms of model calls and elapsed time in seconds. The runtime is averaged over 10 independent runs for two benchmarks. For *LM*, Inconsistent KernelSHAP-IQ never reaches the MSE error levels and Permutation requires more than 2^{14} evaluations.

Benchmark	Error	Metric	KernelSHAP-IQ	Inc. KernelSHAP-IQ	SVARM-IQ	SHAP-IQ	Permutation
CH	*MSE at 2e^{-3}*	*Model Calls*	75	50	130	180	195
		Time (s)	0.023	0.012	0.035	0.033	0.024
	MSE at 1e^{-3}	*Model Calls*	85	70	170	240	600
		Time (s)	0.026	0.017	0.043	0.037	0.071
LM	*MSE at 1e^{-3}*	*Model Calls*	2800	-	2800	5500	>16384
		Time (s)	35.5	-	35.6	95.9	-
	MSE at 5e^{-4}	*Model Calls*	4000	-	4000	7200	>16384
		Time (s)	70.2	-	71.6	126.1	-

D.2. Validations of Higher-Order Conjecture

In the following, we empirically validate Conjecture 3.8 and Conjecture 3.9. We let the numbers of players be $n = 2, \dots, 11$ and the order of interactions $k = 1, \dots, \lfloor n/2 \rfloor$, since the conjectures hold for $n \geq 2k$.

Validation of Conjecture 3.8 We let $\mu_\infty = 10^7$, generate the matrices $\mathbf{X}_k$ and $\mathbf{W}_k$ and compute the inverse $\mathbf{A}_k$ using standard *numpy* functions. We then compare the results with the proposed inverse from Conjecture 3.8. Lastly, we compute the MSE of all elements and assert that this error is less than 10^{-10}.

Validation of Conjecture 3.9 To validate Conjecture 3.9, we randomly generate 10 instances of SOUMs containing 1000 randomly generated interactions, where the size of each interaction is uniformly distributed. We compute the GT SIIs for these games and compare them with the empirically computed scores via Conjecture 3.9. Lastly, we average the MSE over all SOUM instances and assert again that this error is below 10^{-10}.

D.3. Additional Approximation Results

This section contains additional experimental results and evaluations. Figures 6 and 7 contain additional plots to the experiments conducted on the CH and BR regression tasks, respectively. KernelSHAP-IQ and inconsistent KernelSHAP-IQ substantially outperform SVARM-IQ, SHAP-IQ, and permutation sampling, while inconsistent KernelSHAP-IQ does not converge to the GT. Figures 8 and 9 show the estimation qualities for the image-related tasks. On the CNN, KernelSHAP-IQ outperforms all existing baselines and inconsistent KernelSHAP-IQ differs greatly from the GT SII values. Interestingly, on the ViT, SVARM-IQ outperforms KernelSHAP-IQ and inconsistent KernelSHAP-IQ for estimating second order SII scores while both kernel-based estimators retrieve better first and third order SII values. Figure 10 shows additional estimation results on the AC classification dataset. On AC, KernelSHAP-IQ, again, achieves state-of-the-art approximation results. Lastly, Figure 12 shows a detailed view of the experiment conducted on the LM highlight the estimation quality of each order individually. The results on the LM show that KernelSHAP-IQ and SVARM-IQ both retrieve equally good estimates for order 2 while KernelSHAP-IQ slightly outperforms SVARM-IQ on order 1 and 3. Lastly, Figure 11 shows the results on the synthetic SOUMs with higher number of players. In both settings inconsistent KernelSHAP and KernelSHAP clearly outperform SVARM-IQ, SHAP-IQ, and permutation sampling. Notably, in these higher player settings the inconsistent version's drawback of not converging to the GT is yet to materialize.

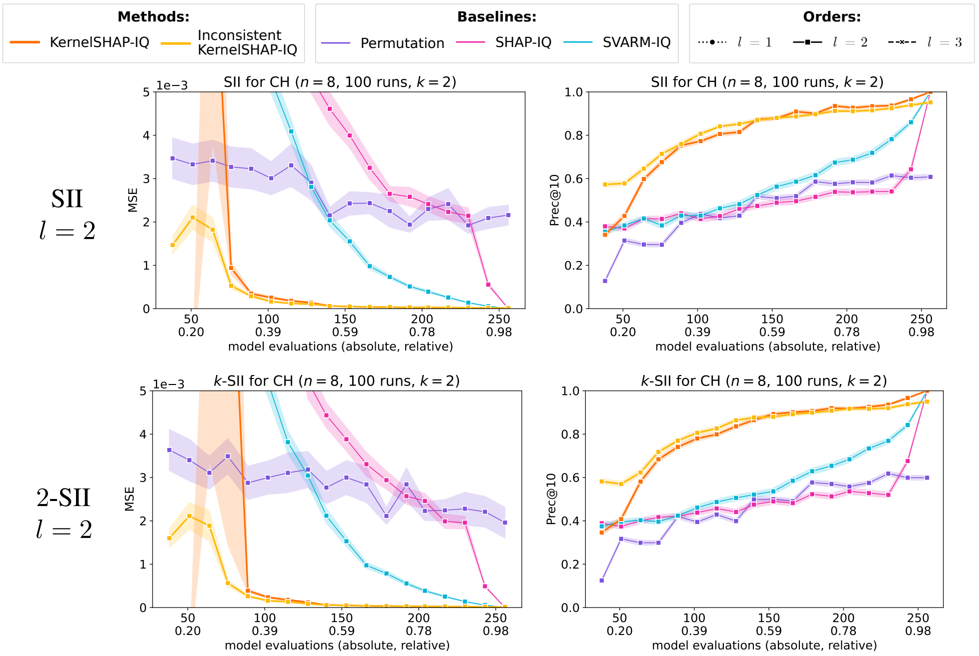

Figure 6. SII (top) and k-SII (bottom) approximation quality in terms of MSE (left) and Prec@10 (right) of the CH dataset.

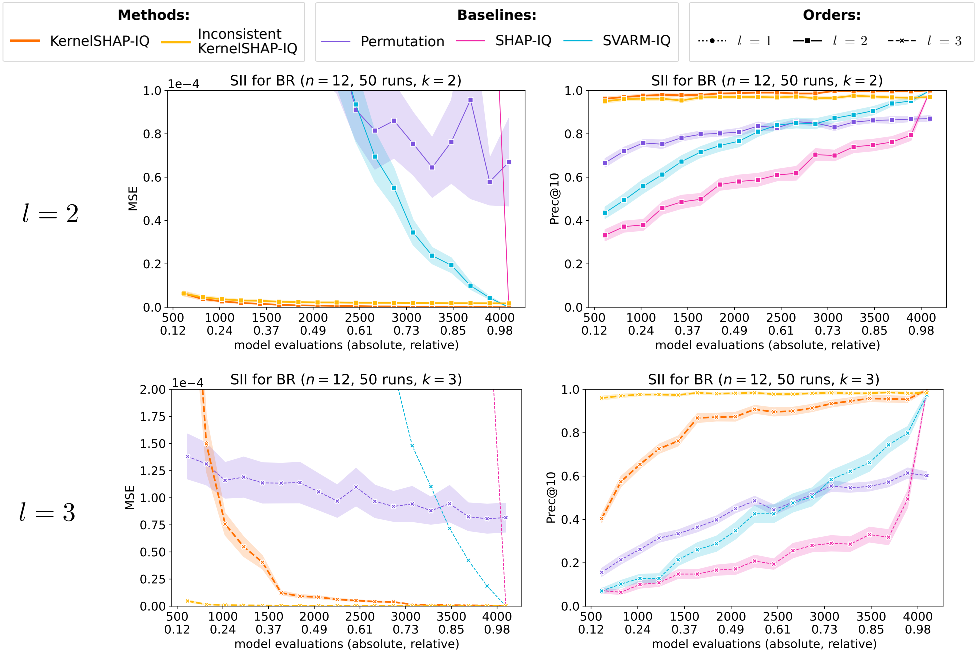

Figure 7. Approximation quality in terms of MSE (left) and Prec@10 (right) of the BR dataset.

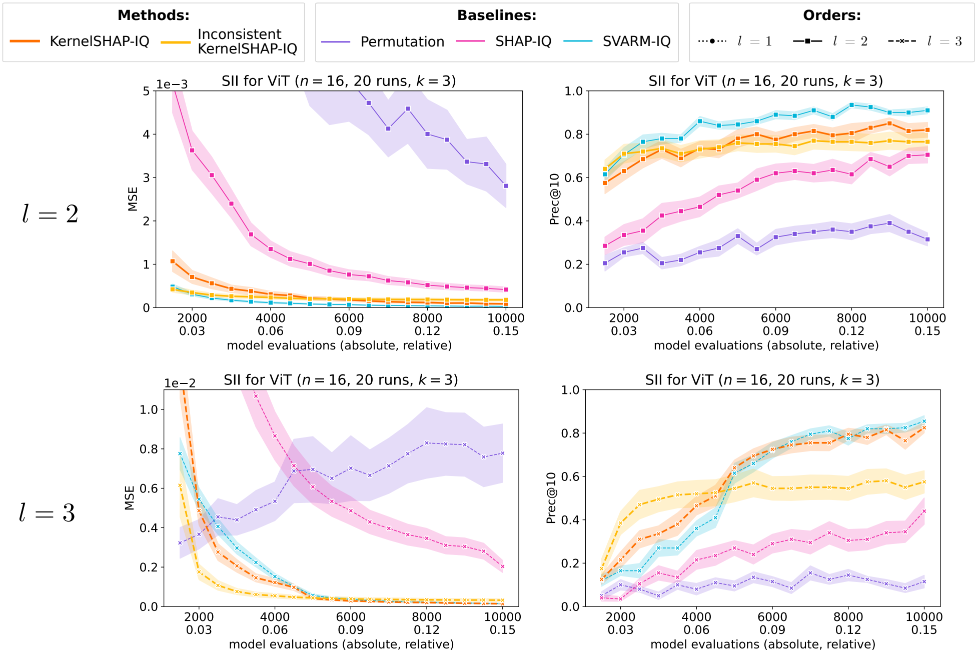

Figure 8. Approximation quality in terms of MSE (left) and Prec@10 (right) of the ViT.

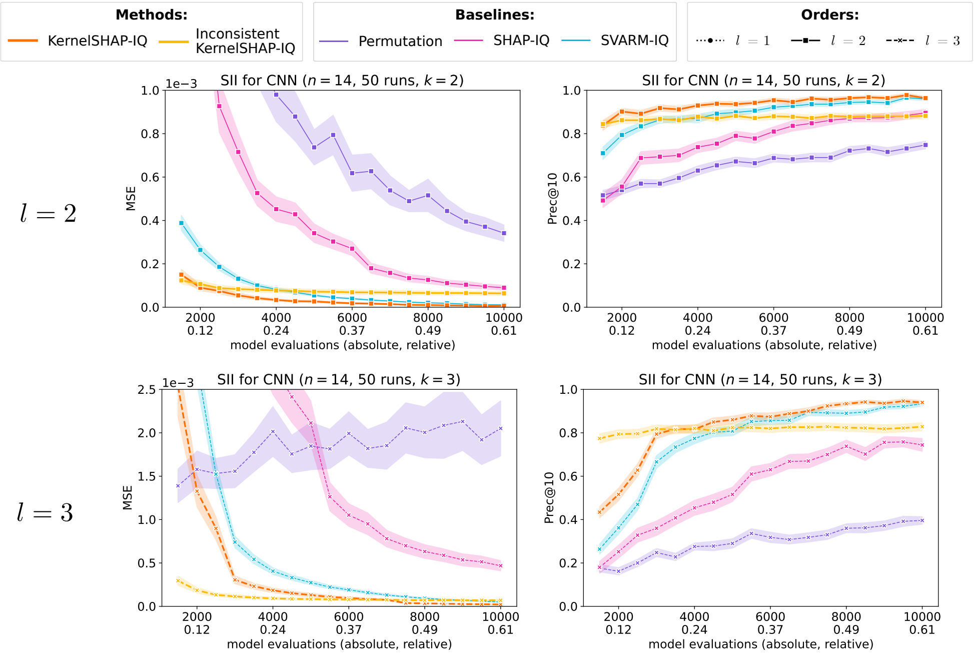

Figure 9. Approximation quality in terms of MSE (left) and Prec@10 (right) of the CNN.

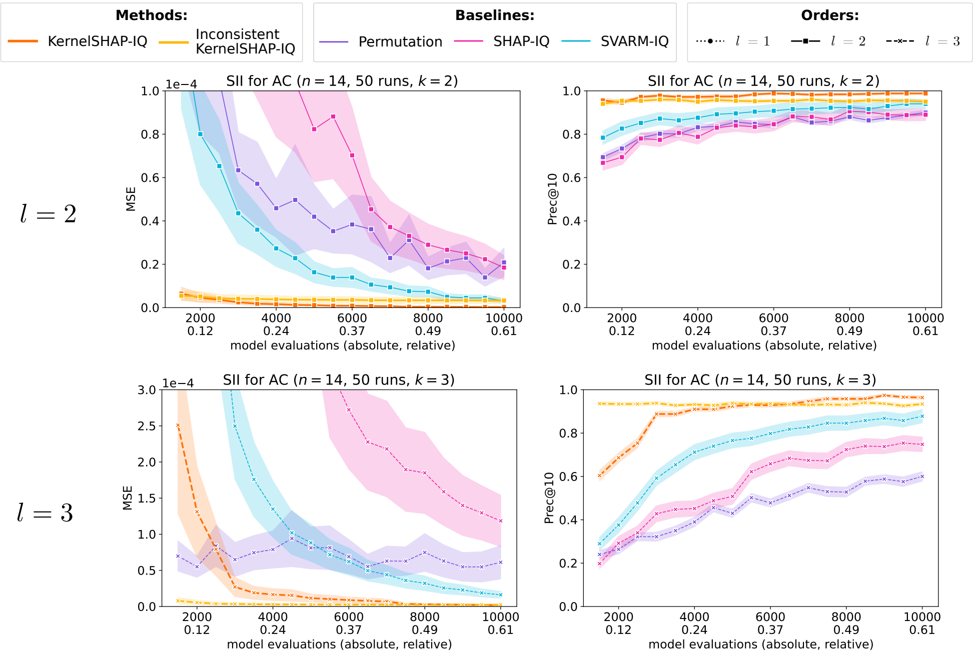

Figure 10. Approximation quality in terms of MSE (left) and Prec@10 (right) of the AC dataset.

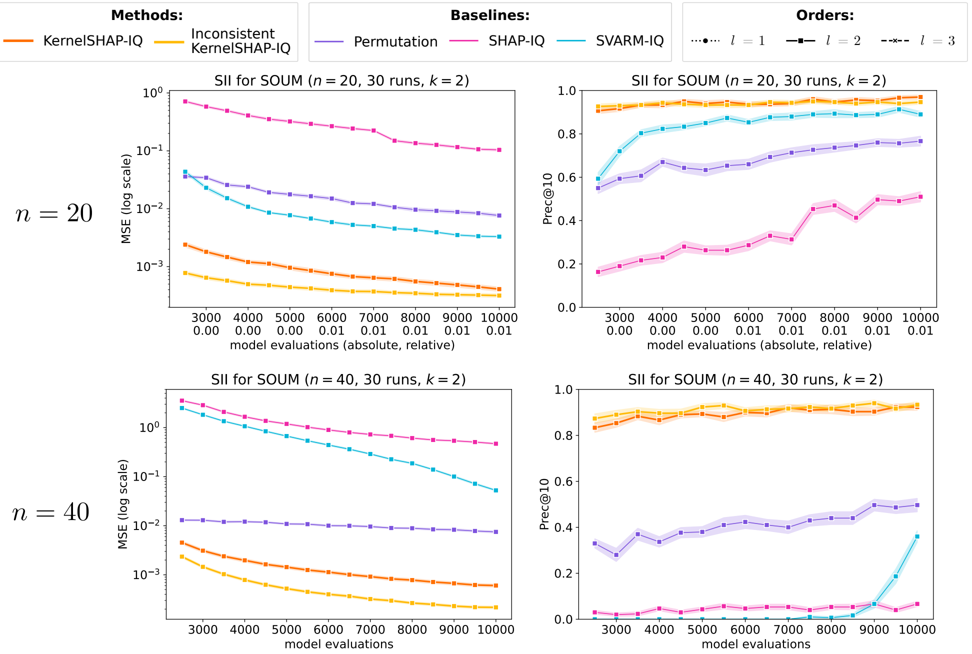

Figure 11. Approximation quality in terms of MSE (left) and Prec@10 (right) for the SOUMs with $n = 20$ (top) and $n = 40$ (bottom).

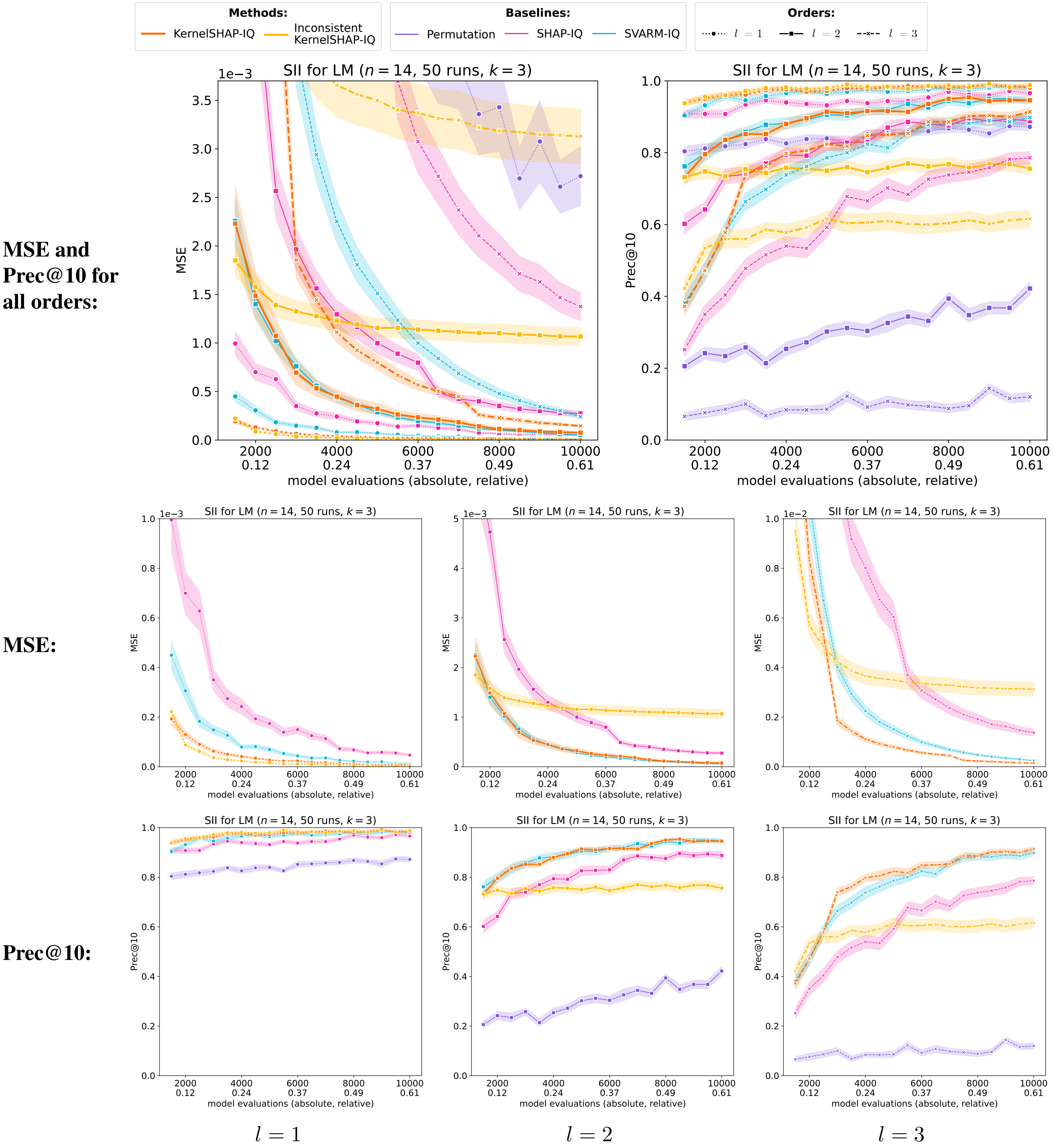

Figure 12. Approximation quality of KernelSHAP-IQ compared to baseline techniques on the LM.