# SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification

**Patrick Kolpaczki**
Paderborn University

**Maximilian Muschalik**
University of Munich (LMU)
Munich Center for Machine Learning

**Fabian Fumagalli**
CITEC
Bielefeld University

**Barbara Hammer**
CITEC
Bielefeld University

**Eyke Hüllermeier**
University of Munich (LMU)
Munich Center for Machine Learning

## Abstract

Addressing the limitations of individual attribution scores via the Shapley value (SV), the field of explainable AI (XAI) has recently explored intricate interactions of features or data points. In particular, extensions of the SV, such as the Shapley Interaction Index (SII), have been proposed as a measure to still benefit from the axiomatic basis of the SV. However, similar to the SV, their exact computation remains computationally prohibitive. Hence, we propose with SVARM-IQ a sampling-based approach to efficiently approximate Shapley-based interaction indices of any order. SVARM-IQ can be applied to a broad class of interaction indices, including the SII, by leveraging a novel stratified representation. We provide non-asymptotic theoretical guarantees on its approximation quality and empirically demonstrate that SVARM-IQ achieves state-of-the-art estimation results in practical XAI scenarios on different model classes and application domains.

## 1 INTRODUCTION

Interpreting black box machine learning (ML) models via feature attribution scores is a widely applied technique in the field of explainable AI (XAI) (Adadi and Berrada, 2018; Covert et al., 2021; Chen et al., 2023). However, in real-world applications, such as genomics (Wright et al., 2016) or tasks involving natural language (Tsang et al., 2020), isolated features are less meaningful. In fact, it was shown that, in the presence of strong feature correlation or higher order interactions, feature attribution scores are not sufficient to capture the reasoning of a trained ML model (Wright et al., 2016; Slack et al., 2020; Sundararajan and Najmi, 2020; Kumar et al., 2020, 2021). As a remedy, *feature interactions* extend feature attributions to arbitrary groups of features (see Figure 1).

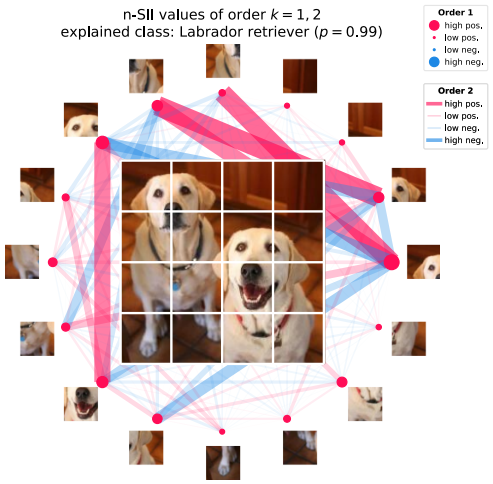



Figure 1: By dividing an ImageNet picture into multiple patches, attribution scores for single patches and interactions scores for pairs aid explaining a vision transformer.

A prevalent approach to define feature attributions is based on the Shapley value (SV) (Shapley, 1953), an axiomatic concept from cooperative game theory

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

that fairly distributes the payout achieved by a group among its members. Extensions of the SV to Shapley-based interaction indices, i.e., interaction indices that reduce to the SV for single players, have been proposed (Grabisch and Roubens, 1999; Bordt and von Luxburg, 2023; Sundararajan et al., 2020; Tsai et al., 2023). Yet, the exact computation of the SV and Shapley-based interactions without further assumptions on the ML model quickly becomes infeasible due to its exponential complexity (Deng and Papadimitriou, 1994).

In this work, we present *SVARM Interaction Quantification* (SVARM-IQ), a novel approximation technique for a broad class of interaction indices, including Shapley-based interactions, which is applicable to any cooperative game. SVARM-IQ extends Stratified SVARM (Kolpaczki et al., 2023) to any-order interactions by introducing a novel representation of interaction indices through stratification.

**Contribution.** Our core contributions include:

1. *SVARM-IQ* (Section 3): A model-agnostic approximation algorithm for estimating Shapley-based interaction scores of any order through leveraging a *stratified* representation.
2. *Theoretical Analysis* (Section 4): We prove, under mild assumptions, that SVARM-IQ is unbiased and provide bounds on the approximation error.
3. *Application* (Section 5): An open-source implementation[1] and empirical evaluation demonstrating SVARM-IQ's superior approximation quality over state-of-the-art techniques.

**Related work.** In cooperative game theory, Shapley-based interactions, as an extension to the SV, were first proposed with the Shapley-Interaction index (SII) (Grabisch and Roubens, 1999). Besides the SII, the Shapley-Taylor Interaction index (STI) (Sundararajan et al., 2020) and Faithful Shapley-Interaction index (FSI) (Tsai et al., 2023) were introduced, which, in contrast to the SII, directly require the efficiency axiom. Beyond Shapley-based interaction indices, extensions of the Banzhaf value were studied by Hammer and Holzman (1992). In ML, limitations of feature attribution scores have been discussed in Wright et al. (2016), Sundararajan and Najmi (2020), and Kumar et al. (2020, 2021) among others. Model-specific interaction measures have been proposed for neural networks (Tsang et al., 2018; Singh et al., 2019; Janizek et al., 2021). Model-agnostic measures were introduced via functional decomposition (Hooker, 2004, 2007) in (Lou et al., 2013; Molnar et al., 2019; Lengerich et al., 2020; Hiabu et al., 2023). Applications include complex language (Murdoch et al., 2018) and image classification (Tsang et al., 2020) models, as well as application domains, such as gene interactions (Wright et al., 2016). Besides pure explanation purposes, e.g. understanding sentiment predictions from NLP models (Fumagalli et al., 2023), Chu and Chan (2020) leveraged the SII to improve feature selection for tree classifiers.

Approximation techniques for the SV have been proposed via permutation sampling (Castro et al., 2009), which has been extended to the SII and STI (Sundararajan et al., 2020; Tsai et al., 2023). For the SV, Castro et al. (2017) demonstrated the impact of stratification on approximation performance. Alternatively, the SV can be represented as a solution to a least squares problem (Charnes et al., 1988), which was exploited for approximation (Lundberg and Lee, 2017; Covert and Lee, 2021) and extended to FSI (Tsai et al., 2023). Recent work proposed a model-agnostic sampling-based approach (Fumagalli et al., 2023) for Shapley-based interactions, which was further linked to Covert and Lee (2021). On the model-specific side Muschalik et al. (2024) extended the polynomial-time exact computation of the SV for local feature importance in decision trees (Lundberg et al., 2020) to the SII. While permutation-based approaches are restricted to update single estimates, Kolpaczki et al. (2023) proposed wit Stratified SVARM a novel approach for the SV that is capable of updating all estimates using only a single value function call.

## 2 SHAPLEY-BASED INTERACTION INDICES

In the following, we are interested in properties of a *cooperative game*, that is a tuple $(\mathcal{N}, \nu)$ containing a *player set* $\mathcal{N} = \{1, \dots, n\}$ with $n \in \mathbb{N}$ players and a *value function* $\nu : 2^{\mathcal{N}} \to \mathbb{R}$ mapping each subset $S \subseteq \mathcal{N}$ of players, also called coalition, to a real-valued number $\nu(S)$. In the field of XAI, the value function typically represents a specific *model behavior* (Covert et al., 2021), such as the prediction of an instance or the dataset loss. The player set represents the entities whose attribution will be determined, e.g., the contribution of features to a prediction or the dataset loss. To determine the worth of individual players, the Shapley value (SV) (Shapley, 1953) can be expressed as a weighted average over marginal contributions.

**Definition 2.1** (Shapley Value (Shapley, 1953))**.** *The SV is*

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

*where* $i \in \mathcal{N}$ *and* $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$.

[1] https://github.com/kolpaczki/svarm-iq

The SV is provably the unique attribution measure that fulfills the following axioms: linearity (linear combinations of value functions yield linear combinations of attribution), dummy (players that do not impact the worth of any coalition receive zero attribution), symmetry (two players contributing equally to all coalitions receive the same attribution), and efficiency (the sum of of all players' attributions equals the worth of all players) (Shapley, 1953). In many ML related applications, however, the attribution via the SV is limited in the presence of strong feature correlation or higher order interaction (Slack et al., 2020; Sundararajan and Najmi, 2020; Kumar et al., 2020, 2021). It is therefore necessary to study *interactions between players* in cooperative games. The SV is a weighted average of marginal contributions $\Delta_i$ of single players, and a natural extension to pairs of players is

$$\Delta_{i,j}(S) := \nu(S \cup \{i,j\}) - \nu(S) - \Delta_i(S) - \Delta_j(S)$$

for $S \subseteq \mathcal{N} \setminus \{i,j\}$. Generalizing this recursion to higher order interactions yields the following definition.

**Definition 2.2** (Discrete Derivative (Fujimoto et al., 2006))**.** *For $K \subseteq \mathcal{N}$, the* K-derivative *of $\nu$ at $S \subseteq \mathcal{N} \setminus K$ is*

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

The Shapley interaction index (SII) was the first axiomatic extension of the SV to higher order interaction (Grabisch and Roubens, 1999). It can be represented as a weighted average of discrete derivatives.

**Definition 2.3** (Shapley Interaction Index (Grabisch and Roubens, 1999))**.** *The SII of $K \subseteq \mathcal{N}$ is defined as*

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

**Cardinal Interaction Indices.** Besides the SII, the Shapley-Taylor interaction index (STI) (Sundararajan et al., 2020) and Faithful Shapley interaction index (FSI) (Tsai et al., 2023) have been proposed as extensions of the SV to interactions. More general, the SII can be viewed as a particular instance of a broad class of interaction indices, known as cardinal interaction indices (CIIs) (Fujimoto et al., 2006), which are defined as a weighted average over discrete derivatives:

$$I_K = \sum_{S \subseteq \mathcal{N} \setminus K} \lambda_{k,|S|} \Delta_K(S)$$

with weights $\lambda_{k,|S|}$. In particular, every interaction index satisfying the (generalized) linearity, symmetry and dummy axiom, e.g., SII, STI and FSI, can be represented as a CII (Grabisch and Roubens, 1999). Beyond Shapley-based interaction indices, CIIs also include other interaction indices, such as a generalized Banzhaf value (Hammer and Holzman, 1992). In Section 3, we propose a unified approximation that applies to any CII. For details about other CIIs and their specific weights, we refer to Appendix B.

The SII is the provably unique interaction index that fulfills the (generalized) linearity, symmetry and dummy axiom, as well as a novel recursive axiom that links higher order interactions to lower order interactions (Grabisch and Roubens, 1999). For interaction indices it is also possible to define a generalized efficiency condition, i.e. that $\sum_{K \subseteq \mathcal{N}, |K| \leq k_{\max}} I_K = \nu(\mathcal{N})$ for a maximum interaction order $k_{\max}$. In ML applications, this condition ensures that the sum of contributions equals the model behavior of $\mathcal{N}$, such as the prediction of an instance. The SII scores can be aggregated to fulfill efficiency, which yield the n-Shapley values (n-SII) (Bordt and von Luxburg, 2023). Furthermore, other variants, such as STI and FSI, extend the SV to interactions by directly requiring an efficiency axiom. In contrast to the SV, however, a unique index is only obtained by imposing further conditions. Similar to the SV, whose computation is NP-hard (Deng and Papadimitriou, 1994), the weighted sum of discrete derivatives requires $2^n$ model evaluations, necessitating approximation techniques.

### 2.1 Approximations of Shapley-based Interaction Scores

Different approximation techniques have been proposed to overcome the computational complexity of Shapley-based interaction indices, which extend on existing techniques for the SV.

**Permutation Sampling.** For the SV, permutation sampling (Castro et al., 2009) was proposed, where the SV is represented as an average over randomly drawn permutations of the player set. For each drawn permutation, the algorithm successively adds players to the subset, starting from the empty set using the given order. By comparing the evaluations successively, the marginal contributions are used to update the estimates. Extensions of permutation sampling have been proposed for the SII (Tsai et al., 2023) and STI (Sundararajan et al., 2020). For the SII, only interactions that appear in a consecutive order in the permutation can be updated, resulting in very few updates per permutation. For the STI, all interaction scores can be updated with a single permutation, however, the computational complexity increases, as the discrete derivatives have to be computed for every subset, resulting

in an increase by a factor of $2^k$ per interaction.

**Kernel-based Approximation.** Besides the weighted average, the SV also admits a representation as a solution to a constrained weighted least square problem (Charnes et al., 1988). This optimization problem requires again $2^n$ model evaluations. However, it was proposed to approximate the optimization problem through sampling and solve the resulting optimization problem explicitly, which is known as KernelSHAP (Lundberg and Lee, 2017). An extension of kernel-based approximation was proposed for FSI (Tsai et al., 2023), but it remains open, whether this approach can be generalized to other indices, while its theoretical properties are unknown.

**Unbiased KernelSHAP and SHAP-IQ.** Unbiased KernelSHAP (Covert and Lee, 2021) constitutes a variant of KernelSHAP to approximate the SV, which yields stronger theoretical results, including an unbiased estimate. While this approach is motivated through a kernel-based approximation, it was shown that it is possible to simplify the calculation to a sampling-based approach (Fumagalli et al., 2023). Using the sampling-based approach, SHAP-IQ (Fumagalli et al., 2023) extends Unbiased KernelSHAP to general interaction indices.

### 2.2 Stratified Approximation for the SV

Stratification partitions a population into distinct subpopulations, known as strata, where sampling is then separately executed for each stratum. If the strata are chosen as homogeneous groups with lower variability, stratified sampling yields a better approximation. First proposed for the SV by Maleki et al. (2013), it was shown empirically that stratification by coalition size can improve the approximation (Castro et al., 2017), while recent work extended it by more sophisticated techniques (Burgess and Chapman, 2021). With Stratified SVARM, Kolpaczki et al. (2023) proposed an approach that abstains from sampling marginal contributions. Instead, it samples coalitions to leverage its novel representation of the SV, which splits the marginal contributions into two coalitions and stratifies them by size. This allows one to assign each sampled coalition to one stratum per player, thus efficiently computing SV estimates for all players simultaneously. Hence in contrast to permutation sampling, Stratified SVARM reaches a new level of efficiency as all estimates are updated using a single model evaluation. In comparison to KernelSHAP, it is well understood theoretically and shows significant performance improvements compared to Unbiased KernelSHAP (Kolpaczki et al., 2023). In the following, we extend Stratified SVARM to Shapley-based interaction indices, and even general CIIs.

## 3 SVARM-IQ: A STRATIFIED APPROACH

Since the practical infeasibility of computing the CII incentivizes its approximation as a remedy, we formally state our considered approximation problem under the fixed-budget setting in Section 3.1. We continue by introducing our stratified representation of the CII in Section 3.2, which stands at the core of our new method *SVARM-IQ* presented in Section 3.3.

### 3.1 Approximation Problem

Given a cooperative game $(\mathcal{N}, \nu)$, an order $k \geq 2$, a budget $B \in \mathbb{N}$, and the weights $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$, with $\mathcal{L}_k := \{0, \ldots, n-k\}$ specifying the desired CII, the goal is to approximate all the latent but unknown CII $I_K$ with $K \in \mathcal{N}_k := \{S \subseteq \mathcal{N} \mid |S| = k\}$ precisely. The budget $B$ is the number of coalition evaluations or in other words accesses to $\nu$ that the approximation algorithm is allowed to perform. It captures a time or resource constraint on the computation and is justified by the fact that the access to $\nu$ frequently imposes a bottleneck on the runtime due to costly inference, manipulation of data, or even retraining of models. We denote by $\hat{I}_K$ the algorithm's estimate of $I_K$. Since we consider randomized algorithms, returning stochastic estimates, the approximation quality of an estimate $\hat{I}_K$ is judged by the following two commonly used measures that are to be minimized: First, the mean squared error (MSE) of any set $K$: $\mathbb{E}\left[(\hat{I}_K - I_K)^2\right]$, and second, a bound on the probability $\mathbb{P}(|\hat{I}_K - I_K| \geq \varepsilon) \leq \delta$ to exceed a threshold $\varepsilon > 0$, commonly known as a $(\epsilon, \delta)$-approximation.

### 3.2 Stratified Representation

Our sampling-based approximation algorithm SVARM-IQ leverages a novel stratified representation of the CII. For the remainder, we stick to the general notion of the CII of any fixed order $k \geq 2$. The concrete interaction type to be approximated can be specified by the weights $\lambda_{k,\ell}$. We stratify the CII $I_K$ by coalition size and split the discrete derivatives $\Delta_K(S)$ into multiple strata to obtain:

$$I_K = \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot I_{K,\ell}^W.$$

with strata terms for all $W \subseteq K$ and $\ell \in \mathcal{L}_k$:

$$I_{K,\ell}^W := \frac{1}{\binom{n-k}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W). \tag{1}$$

This representation is a generalization of the SV representation utilized by Stratified SVARM (Kolpaczki et al., 2023) as it extends from the SV to the CII. Since each stratum contains $\binom{n-k}{\ell}$ many coalitions, $I^W_{K,\ell}$ is a uniform average of all eligible coalition worths and hence we obtain its estimate $\hat{I}^W_{K,\ell}$ by taking the sample-mean of evaluated coalitions belonging to that particular stratum. Further, we can express any CII by means of the strata $I^W_{K,\ell}$ trough manipulating their weighting according to the weights $\lambda_{k,\ell}$. Subsequently, the aggregation of the strata estimates, mimicking our representation, yields the desired CII estimate:

$$\hat{I}_K = \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \hat{I}^W_{K,\ell}.$$

Further, we demonstrate the popular special case of SII between pairs, i.e., $k = 2$, in Appendix D.

### 3.3 SVARM-IQ

Instead of naively sampling coalitions separately from each of the $2^k\binom{n}{k}(n-k+1)$ many strata, we propose with SVARM-IQ a more sophisticated mechanism, similar to Kolpaczki et al. (2023), which leverages the stratified representation of the CII.

**Update Mechanism.** SVARM-IQ (given in Algorithm 1 and Figure 2) updates for a single sampled coalition $A \subseteq \mathcal{N}$ one strata estimate of each of the $\binom{n}{k}$ many considered subsets $K$. This is made feasible by the observation that any coalition $A$ belongs into exactly one stratum associated with $I_K$. This is in the spirit of the *maximum sample reuse* principle, employed previously for the Banzhaf value (Wang and Jia, 2023) and the SV (Kolpaczki et al., 2023) with the underlying motivation that each seen observation should be utilized to update all interaction estimates. To be more precise, for each $K \in \mathcal{N}_k$ we update

$$\hat{I}^W_{K,\ell} \text{ with } W = A \cap K \text{ and } \ell = |A| - |W|. \quad (2)$$

Notably, our sampling ensures that for *every interaction* $A \setminus K \sim \text{unif}(\{S \subseteq \mathcal{N} \setminus K \mid |S| = \ell\})$, as the probability of $A \setminus K$ conditioned on $W = A \cap K$ and $\ell = |A| - |A \cap K|$ is uniform. This is required as $\hat{I}^W_{K,\ell}$ is a uniform average, cf. Eq. (1), and allows to update estimates for every interaction by sampling a single subset $A$. Considering the limited budget $B$, this update rule elicits information from $\nu$ in a more "budget-efficient" manner, since it contributes to $\binom{n}{k}$ many estimates with only a single evaluation. To guide the sampling, we first draw in each time step $b$ a coalition size $a_b$ from a probability distribution $P_k$ over the eligible sizes, and draw then $A_b$ uniformly at random among all coalitions of size $a_b$. We store the evaluated worth $\nu(A_b)$ in order to reuse it for all estimate updates, one for each $K$. This is done by calling `UpdateMean` (see Appendix C), which sets the associated estimate $\hat{I}^W_{K,\ell}$ to the new average, taking the sampled worth $\nu(A_b)$ and the number of so far observed samples $c^W_{K,\ell}$ of that particular estimate into account. We set $P_k$ to be the uniform distribution over all sizes, i.e., $P_k = \text{unif}(0, n)$. A specifically tailored distribution for $k = 2$ allows us to express sharper theoretical results in Section 4.

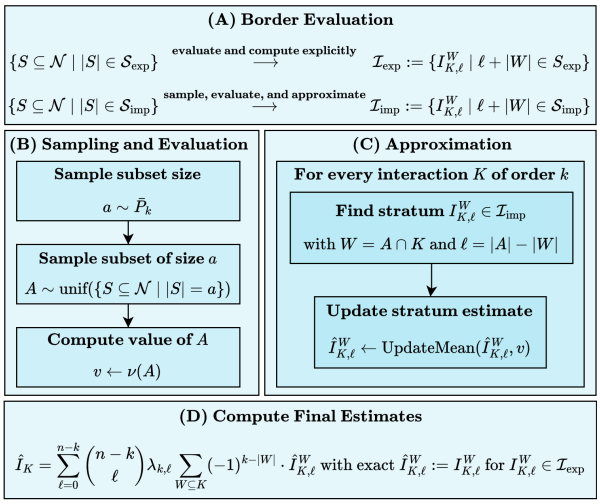



Figure 2: Schematic overview of SVARM-IQ.

**Border Sizes.** Further, we enhance our approach by transferring a technique, introduced by Fumagalli et al. (2023). We observe that for very low and very high $s$ only a few coalitions of size $s$ exist, $\binom{n}{s}$ many to be precise. Thus, evaluating all these coalitions and calculating the associated strata $I^W_{K,\ell}$ explicitly upfront saves budget, as it avoids duplicates, i.e., coalitions sampled multiple times. Given the budget $B$ and the probability distribution over sizes $P_k$, we determine a set of subset sizes $\mathcal{S}_{\text{exp}} = \{0, \ldots, s_{\text{exp}}, n - s_{\text{exp}}, \ldots, n\}$, for which the expected number of samples exceeds the number of coalitions of each subset size. Consequently, we evaluate *all coalitions* of sizes in $\mathcal{S}_{\text{exp}}$, i.e., $S \subseteq \mathcal{N}$ with $|S| \in \mathcal{S}_{\text{exp}}$. From the remaining sizes $\mathcal{S}_{\text{imp}} := \{s_{\text{exp}} + 1, \ldots, n - s_{\text{exp}} - 1\}$, we sample coalitions. This split allows to compute all strata

$$\mathcal{I}_{\text{exp}} := \{I^W_{K,\ell} \mid \ell + |W| \in S_{\text{exp}}\}$$

*explicitly*, which follows from Eq. (2) and $\ell + |W| = |A|$. The remaining strata

$$\mathcal{I}_{\text{imp}} := \{I^W_{K,\ell} \mid \ell + |W| \in \mathcal{S}_{\text{imp}}\}$$

are *approximated* with $\hat{I}^W_{K,\ell}$ by sampling coalitions. The procedure to determine $\mathcal{S}_{\text{exp}}$ and $\mathcal{S}_{\text{imp}}$, named `ComputeBorders` (see Appendix C), is applied before

**Algorithm 1** SVARM-IQ

**Input:** $(\mathcal{N}, \nu)$, $B \in \mathbb{N}$, $k \in \{1, \dots, n\}$, $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$
$\hat{I}^W_{K,\ell}, c^W_{K,\ell} \leftarrow 0 \; \forall K \in \mathcal{N}_k, \ell \in \mathcal{L}_k, W \subseteq K$
$\mathcal{S}_{\text{exp}}, \mathcal{S}_{\text{imp}} \leftarrow$ ComputeBorders
$\bar{B} \leftarrow B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s}$
**for** $b = 1, \dots, \bar{B}$ **do**
  Draw size $a_b \in \mathcal{S}_{\text{imp}} \sim \bar{P}_k$
  Draw $A_b$ from $\{S \subseteq \mathcal{N} \mid |S| = a_b\}$ u.a.r.
  $v_b \leftarrow \nu(A_b)$ ▷ store coalition worth
  **for** $K \in \mathcal{N}_k$ **do**
    $W \leftarrow A_b \cap K$ ▷ get stratum set
    $\ell \leftarrow a_b - |W|$ ▷ get stratum size
    $\hat{I}^W_{K,\ell} \leftarrow$ UpdateMean$(\hat{I}^W_{K,\ell}, c^W_{K,\ell}, v_b)$
    $c^W_{K,\ell} \leftarrow c^W_{K,\ell} + 1$ ▷ increment counter
  **end for**
**end for**
$\hat{I}_k \leftarrow \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \hat{I}^W_{K,\ell} \; \forall K \in \mathcal{N}_k$
**Output:** $\hat{I}_K$ for all $K \in \mathcal{N}_k$

the sampling loop in Algorithm 1. Hence, SVARM-IQ enters its sampling loop with a leftover budget of $\bar{B} := B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s}$, and repeatedly applies the update mechanism. The distribution $P_k$ is altered to $\bar{P}_k$ by setting $\bar{P}_k(s) = 0$ for all $s \in \mathcal{S}_{\text{exp}}$ and upscaling all entries $s \in \mathcal{S}_{\text{imp}}$ such that they sum up to 1. Note that this technique yields exact CII values for $B = 2^n$.

**Approximating Multiple Orders and Indices.** SVARM-IQ is not restricted to approximate only one specific order $k$ at the time. Quite to the contrary, it can be extended to maintain strata estimates $\hat{I}^W_{K,\ell}$ for multiple orders, which are then simultaneously updated within the sampling loop without imposing further budget costs. The aggregation to interaction estimates $\hat{I}_K$ is then carried out for each considered subset $K$ separately. Note that this also entails the SV, i.e., $k = 1$, thus allowing one to approximate attribution and interaction simultaneously. Since the stratification allows to combine the strata to any CII, SVARM-IQ can approximate multiple CII's at the same time, notably without even the need to specify them during sampling. This can be realized by specifying multiple weighting sequences $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$, one for each CII of interest, and performing the final estimate computation $\hat{I}_K$ for each type. Note that this comes without incurring any additional budget cost.

## 4 THEORETICAL RESULTS

In the following, we present the results of our theoretical analysis for SVARM-IQ. All proofs are deferred to Appendix E. In order to make the analysis feasible, a natural assumption is to observe at least one sample for each approximated stratum $I^W_{K,\ell} \in \mathcal{I}_{\text{imp}}$. We realize this requirement algorithmically only for the remainder of this chapter by executing a WarmUp procedure (see Appendix 3) between ComputeBorders and SVARM-IQ's sampling loop. For each $I^W_{K,\ell} \in \mathcal{I}_{\text{imp}}$ it samples a coalition $A \subseteq \mathcal{N} \setminus K$ of size $\ell$ and sets $\hat{I}^W_{K,\ell}$ to $\nu(A \cup W)$. Hence, SVARM-IQ enters its sampling loop with a leftover budget of $\tilde{B} := B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s} - |\mathcal{I}_{\text{imp}}|$. We automatically set $s_{\text{exp}} \geq 1$, which consumes only $2n + 2$ evaluations. Hence for $n = 3$, all strata are already explicitly calculated. Since ComputeBorders evaluates then at least all coalitions of size $s \in \{0, 1, n-1, n\}$, the initial distribution $P_k$ over sizes has support $\{2, \dots, n-2\}$. For $k \geq 3$, this allows us to specify $P_k$ to be the uniform distribution:

$$P_k(s) := \frac{1}{n-3} \text{ for all } s \in \{2, \dots, n-2\}.$$

Further for the remainder of the analysis, we use a specifically tailored distribution in the case of $k = 2$:

$$P_2(s) := \begin{cases} \frac{\beta_n}{s(s-1)} & \text{if } s \leq \frac{n-1}{2} \\ \frac{\beta_n}{(n-s)(n-s-1)} & \text{if } s \geq \frac{n}{2} \end{cases}$$

with $\beta_n = \frac{n^2-2n}{2(n^2-4n+2)}$ for even $n \geq 4$ and $\beta_n = \frac{n-1}{2(n-3)}$ for odd $n \geq 5$. This allows us to express sharper bounds in comparison to the uniform distribution.

**Notation and assumptions.** We introduce some notation, coming in helpful in expressing our results legibly. For any $w \in \{0, \dots, k\}$ we denote by $\mathcal{L}^w_k := \{\ell \in \mathcal{L}_k \mid \ell + w \in \mathcal{S}_{\text{imp}}\}$. For any $K \in \mathcal{N}_k$ and $\ell \in \mathcal{L}_k$ let $A_{K,\ell}$ be a random coalition with distribution $\mathbb{P}(A_{K,\ell} = S) = \binom{n-k}{\ell}^{-1}$ for all $S \subseteq \mathcal{N} \setminus K$ with $|S| = \ell$. For any $W \subseteq K$ we denote the *stratum variance* by $\sigma^2_{K,\ell,W} := \mathbb{V}[\nu(A_{K,\ell} \cup W)]$ and the *stratum range* by $r_{K,\ell,W} := \max_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W) - \min_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W)$. For a comprehensive overview of the used notation, we refer to Appendix A. As our only assumptions, we demand $n \geq 4$ and the budget to be large enough to execute ComputeBorders, WarmUp , and the sampling loop for one iteration, i.e., $\tilde{B} > 0$.

**Unbiasedness, Variance, and MSE.** We begin by showing that SVARM-IQ's estimates are unbiased, which is not only desirable but will also turn out useful shortly after in our analysis.

**Theorem 4.1.** *SVARM-IQ's CII estimates are unbiased for all $K \in \mathcal{N}_k$, i.e., $\mathbb{E}[\hat{I}_K] = I_K$.*

The unbiasedness enables us to reduce the MSE of any $\hat{I}_K$ to its variance. In fact, the bias-variance decomposition states that $\mathbb{E}[(\hat{I}_K - I_K)^2] = (\mathbb{E}[\hat{I}_K] - I_K)^2 +$

$\mathbb{V}[\hat{I}_K]$. Hence, a variance analysis of the obtained estimates suffices to bound the MSE. The variance of $\hat{I}_K$ is tightly linked to the number of samples SVARM-IQ collects for each stratum estimate $\hat{I}^W_{K,\ell}$. At this point, we distinguish in our analysis between $k = 2$ and $k \geq 3$ to obtain sharper bounds for the former case facilitated by our carefully designed probability distribution $P_2$ over coalition sizes. To keep the presented results concise, we introduce $\gamma_k := 2(n-1)^2$ for $k = 2$ and $\gamma_k := n^{k-1}(n-k+1)^2$ for all $k \geq 3$. This stems from the aforementioned difference in precision on the lower bound of collected samples.

**Theorem 4.2.** *For any $K \in \mathcal{N}_k$ the variance of the CII estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by*

$$\mathbb{V}\left[\hat{I}_K\right] \leq \frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W}.$$

Note that our efforts in optimizing the analysis for $k = 2$ reduced the bound by a factor of $\frac{n}{2}$ in comparison to substituting $k$ with 2 in our bound for the general case. This is caused by the severe increase in complexity when trying to give a lower bound for the number of samples each stratum receives. Although our approach allows one to obtain a sharper bound for special cases as $k = 3$ or $k = 4$ with a similarly dedicated analysis, we abstain from doing so as we prioritize a concise presentation of our results.

**Corollary 4.3.** *For any $K \in \mathcal{N}_k$, the MSE of $\hat{I}_K$ returned by SVARM-IQ is bounded by $\mathbb{E}[(\hat{I}_K - I_K)^2] \leq$*

$$\frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W}.$$

We state this result more explicitly for the frequently considered interaction type: the SII for pairs of players $i$ and $j$. In this case our bound boils down to

$$\mathbb{E}\left[\left(\hat{I}^{\text{SII}}_{i,j} - I^{\text{SII}}_{i,j}\right)^2\right] \leq \frac{2}{\tilde{B}} \sum_{W \subseteq \{i,j\}} \sum_{\ell \in \mathcal{L}_2^{|W|}} \sigma^2_{i,j,\ell,W}.$$

The simplicity achieved by this result supports a straightforward and natural interpretation. The MSE bound of each SII estimate is inversely proportional to the available budget for the sampling loop and each stratum variance contributes equally to its growth.

We *intentionally abstain* from expressing our bounds in asymptotic notation w.r.t. $B$ and $n$ only, as it would not do justice to the motivation behind employing stratification. The performance of SVARM-IQ is based on lower strata variances (and also strata ranges) compared to the whole population of all coalition values within the powerset of $\mathcal{N}$. This improvement can not be reflected adequately by the asymptotics in which the variances vanish to constants.

**$(\epsilon, \delta)$-Approximation.** Combining Theorem 4.2 with Chebyshev's inequality immediately yields a bound on the probability that the absolute error of a fixed $\hat{I}_K$ exceeds some $\varepsilon > 0$ given the budget at hand.

**Corollary 4.4.** *For any $K \in \mathcal{N}_k$, the absolute error of $\hat{I}_K$ returned by SVARM-IQ exceeds some fixed $\varepsilon$ with probability of at most $\mathbb{P}(|\hat{I}_K - I_K| \geq \varepsilon) \leq$*

$$\frac{\gamma_k}{\varepsilon^2 \tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W}.$$

One can easily rearrange the terms to find the minimum budget required to obtain $\mathbb{P}(|\hat{I}_K - I_K| \leq \varepsilon) \geq 1 - \delta$ for a given $\delta > 0$. Note that this bound still depends on the unknown strata variances. Further, we provide another bound in Theorem 4.5 (see Appendix E.4), resulting from a slightly more laborious usage of Hoeffding's inequality which takes the strata ranges into account. To the best of our knowledge there exists no theoretical analysis for permutation sampling of CIIs. SHAP-IQ is like wise unbiased, but its theoretical analysis (Theorem 4.3) Fumagalli et al. (2023) does not provide such detail for fixed $n$ and $k$.

# 5 EXPERIMENTS

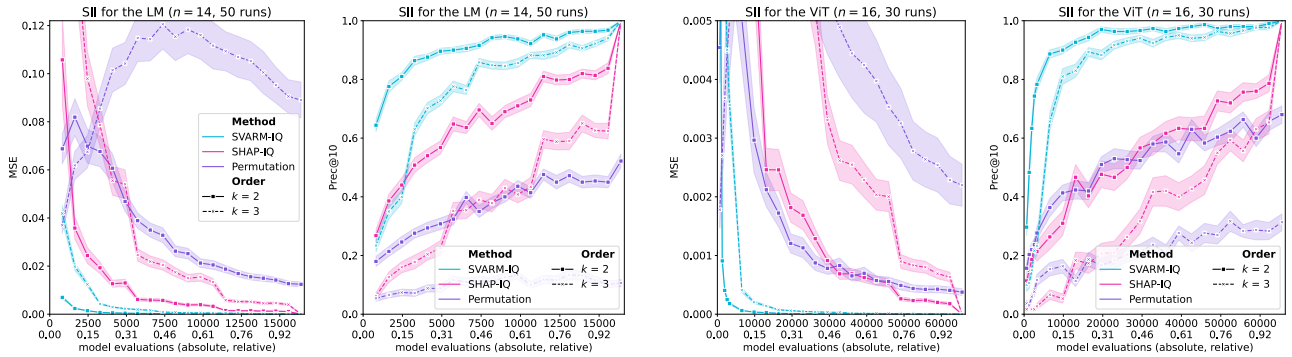
We empirically evaluate SVARM-IQ's approximation quality in different XAI application scenarios and compare it with current state-of-the-art baselines.

**Baselines.** In the case of estimating SII and STI scores, we compare SVARM-IQ to SHAP-IQ (Fumagalli et al., 2023) and permutation sampling (Sundararajan et al., 2020; Tsai et al., 2023). For FSI, we compare against the kernel-based regression approach (Tsai et al., 2023) instead of permutation sampling.

Table 1: Overview of the XAI tasks and models used

| Task | Model ID | Removal Strategy | $n$ | $\mathcal{Y}$ |
|---|---|---|---|---|
| LM | DistilBert | Token Removal | 14 | $[-1, 1]$ |
| ViT | ViT-32-384 | Token Removal | 9,16 | $[0, 1]$ |
| CNN | ResNet18 | Superpixel Marginalization | 14 | $[0, 1]$ |

**Explanation Tasks.** Similar to Fumagalli et al. (2023) and Tsai et al. (2023), we evaluate the approximation algorithms based on different real-world ML models and classical XAI scenarios (cf. Table 1). First, we compute interaction scores to explain a sentiment analysis *language model* (LM), which is a fine-tuned version of `DistilBert` (Sanh et al., 2019) on the




(a) MSE and Prec@10 for the LM

(b) MSE and Prec@10 for the ViT

Figure 3: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines for estimating order $k = 2, 3$ SII on the LM (a; $n = 14$) and the ViT (b; $n = 16$). Shaded bands represent the standard error over 50, respectively 30 runs.

IMDB (Maas et al., 2011) dataset. Second, we investigate two types of image classification models, which were pre-trained on ImageNet (Deng et al., 2009). We explain a *vision transformer* (ViT), (Dosovitskiy et al., 2021), and a `ResNet18` *convolutional neural network* (CNN) (He et al., 2016a). The ViT operates on patches of 32 times 32 pixels and is abbreviated with `ViT-32-384`. The `torch` versions of the LM, ViT, and the CNN are retrieved from Wolf et al. (2020) and Paszke et al. (2017). For further descriptions on the models and feature removal strategies aligned with Covert et al. (2021), we refer to Appendix F.

**Measuring Performance.** To assess the performance of the different approximation algorithms, we measure the mean squared error averaged over all $K \in \mathcal{N}_k$ (MSE; lower is better) and the precision at ten (Prec@10; higher is better) of the estimated interaction scores compared to pre-computed ground-truth values (GTV). Prec@10 measures the ratio of correctly identifying the ten highest (absolute) interaction values. The GTV for each run are computed exhaustively with $2^n$ queries to the black box models. All results are averaged over multiple independent runs.

**Approximation Quality for SII.** We compare SVARM-IQ against permutation sampling and SHAP-IQ at the LM and ViT explanation tasks for approximating all SII values of order $k = 2$ and $k = 3$ in Figure 3. Across both considered measures, MSE and Prec@10, SVARM-IQ demonstrates superior approximation quality. Noteworthy is SVARM-IQ's steep increase in approximation quality in the earlier budget range allowing applications with limited computational resources. Based on our theoretical findings, we assume the stratification by size in combination with the splitting of discrete derivatives to be the cause for

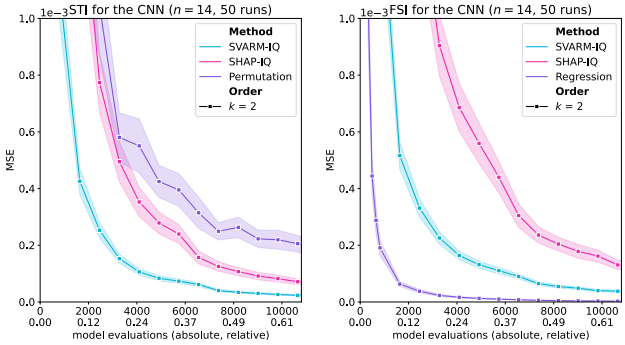



Figure 4: Comparison of SVARM-IQ and baselines for STI (left) and FSI (right) on the CNN. Shaded bands represent the standard error over 50 runs.

the observed behavior. Most plausibly coalitions of the same size and sharing a predetermined set, as encompassed by each stratum $I^W_{K,\ell}$, vary less in their worth than the whole population of coalitions. Consequently, the associated variance $\sigma^2_{K,\ell,W}$ is considerably lower, leading to faster convergence of the estimate $\hat{I}^W_{K,\ell}$.

**Example Use-Case of n-SII Values.** Precise estimates allow to construct high-quality n-SII scores as proposed by Bordt and von Luxburg (2023). Figure 1 illustrates how n-SII scores can be used to explain the ViT with 16 patches for an image of two correctly classified Labradors. All individual patches receive positive attribution scores ($k = 1$) of varying degree, leading practitioners to assume that patches with similar attribution are of equal importance. However, enhancing the explanation with second order interactions ($k = 2$), reveals how the interplay between patches containing complementing facial parts, like the eyes and the mouth, strongly influences the model's prediction towards the correct class label. On the contrary,

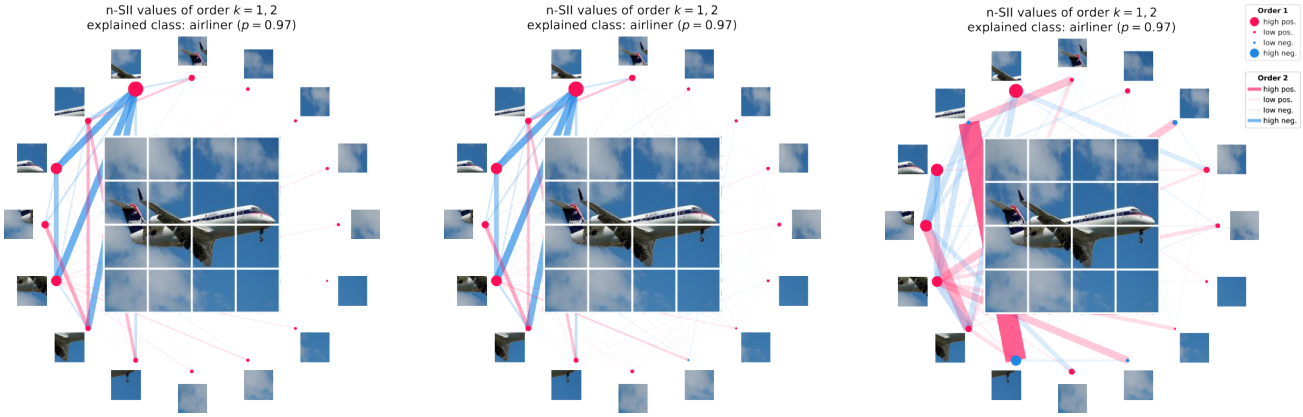



Figure 5: Comparison of ground-truth n-SII values of order $k = 1$ and $k = 2$ for the predicted class probability of a ViT for an ImageNet picture sliced into a grid of $n = 16$ patches (left) against n-SII values estimated by SVARM-IQ (center) and permutation sampling (right). The exact computation requires 65,536 model evaluations while the budget of both approximators is limited by 5000, making up only 7.6% of the space to sample.

tiles depicting the same parts, e.g. those containing eyes, show negative interaction, allowing to conclude that the addition of one in the presence of the other is on average far less impactful than their individual contribution. Solely observing the monotony of the individual scores would have arguably led to overlook this insight. We describe this further in Appendix G.2.

**Estimating FSI and STI.** Further, we compute different CIIs of a fixed order with SVARM-IQ and consistently achieve high approximation quality. We summarize the results on the CNN in Figure 4. For STI, SVARM-IQ, again, outperforms both sampling-based baselines. The kernel-based regression estimator, which is only applicable to the FSI index, yields lower approximation errors than SVARM-IQ. Similar to SV estimation through KernelSHAP (Lundberg and Lee, 2017), this highlights the expressive power of the least-squares representation available for FSI.

**Instance-wise comparison.** Lastly, we compare in Figure 5 SVARM-IQ's n-SII estimates of order $k = 1$ and $k = 2$ against those of permutation sampling and the ground truth for single a single instance. The ground truth interaction is computed upfront for the predicted class probability of the ViT for a specified image sliced into a grid of 16 patches, and both approximation algorithms are executed for a single run with a budget of 5000 model evaluations, thus consuming only 7.6% of the budget necessary to compute GTV exactly. The estimates obtained by SVARM-IQ show barely any visible difference to the human eye. In fact, SVARM-IQ's approximation replicates the ground truth with only a fraction of the number of model evaluations that are necessary for its exact computation. Hence, it significantly lowers the computational burden for precise explanations. On the contrary, permutation sampling yields estimated importance and interaction scores which are afflicted with evident imprecision. This lack in approximation quality has the potential to cause misguiding explanations. More comparisons are shown in Appendix G.2.

## 6 CONCLUSION

We proposed SVARM-IQ, a new sampling-based approximation algorithm for interaction indices based on a stratified representation to maximize budget efficiency. SVARM-IQ is capable of approximating all types and orders of cardinal interactions simultaneously, including the popular SII. Consequently, as the special case of SVs is also entailed, this facilitates the approximation of feature importance and interaction simultaneously, thus offering an enriched explanation. Besides proving theoretical results, we empirically demonstrated SVARM-IQ's advantage against current state-of-the-art baselines. Its model-agnostic nature and domain-independence allow practitioners to obtain high-quality interaction scores for various entity types such as features or data points.

**Limitations and Future Work.** Due to SVARM-IQ's stratification, the number of maintained strata estimates grows exponentially with the interaction order $k$. This space complexity poses a challenge for large interaction orders. As a pragmatic remedy, future work may consider the approximation of interaction scores for a smaller number of sets or a coarser stratification by size. Lastly, it still remains unclear whether the performance of the kernel-based regression estimator available for FSI and the SV can be transferred to other types of CIIs like the SII or STI indices.

## Acknowledgements

This reserach was supported supported by the research training group Dataninja (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia. Maximilan Muschalik and Fabian Fumagalli gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

## References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282.

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Bordt, S. and von Luxburg, U. (2023). From Shapley Values to Generalized Additive Models and back. In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 709–745. PMLR.

Burgess, M. A. and Chapman, A. C. (2021). Approximating the shapley value using stratified empirical bernstein sampling. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, (IJCAI) 2021*, pages 73–81. ijcai.org.

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

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

Chao, M. T. and Strawderman, W. E. (1972). Negative Moments of Positive Random Variables. *Journal of the American Statistical Association*, 67(338):429–431.

Charnes, A., Golany, B., Keane, M., and Rousseau, J. (1988). *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, volume 11 of *Advanced Studies in Theoretical and Applied Econometrics*, page 123–133. Springer Netherlands.

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

Chu, C. and Chan, D. P. K. (2020). Feature selection using approximated high-order interaction components of the shapley value for boosted tree classifier. *IEEE Access*, 8:112742–112750.

Covert, I. and Lee, S. (2021). Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, volume 130 of *Proceedings of Machine Learning Research*, pages 3457–3465. PMLR.

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

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

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations (ICLR 2021)*. OpenReview.net.

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

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. (2023). SHAP-IQ: Unified Approximation of any-order Shapley Interactions. *CoRR*, abs/2303.01179.

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

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

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

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

Hiabu, M., Meyer, J. T., and Wright, M. N. (2023). Unifying local and global model explanations by functional decomposition of low dimensional structures. In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 7040–7060. PMLR.

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

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

Janizek, J. D., Sturmfels, P., and Lee, S. (2021). Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *Journal of Machine Learning Research*, 22:104:1–104:54.

Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. (2023). Approximating the Shapley Value without Marginal Contributions. *CoRR*, abs/2302.00736.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. A. (2021). Shapley Residuals: Quantifying the limits of the Shapley value for explanations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021 (NeurIPS 2021)*, pages 26598–26608.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR.

Lengerich, B. J., Tan, S., Chang, C., Hooker, G., and Caruana, R. (2020). Purifying Interaction Effects with the Functional ANOVA: An Efficient Algorithm for Recovering Identifiable Additive Models. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2402–2412. PMLR.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013)*, pages 623–631. ACM.

Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NeurIPS 2017)*, pages 4765–4774.

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

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

Molnar, C., Casalicchio, G., and Bischl, B. (2019). Quantifying Model Complexity via Functional Decomposition for Better Post-hoc Interpretability. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*, volume Communications in Computer and Information Science, pages 193–204. Springer, Cham.

Murdoch, W. J., Liu, P. J., and Yu, B. (2018). Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *6th International Conference on Learning Representations (ICLR 2018)*.

Muschalik, M., Fumagalli, F., , Hüllermeier, E., and Hammer, B. (2024). Beyond TreeSHAP: Efficient Computation of Any-Order Shapley Interactions for Tree Ensembles. *CoRR*, abs/2401.12069.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *Workshop at Conference on Neural Information Processing Systems (NeurIPS 2017)*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*.

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

Singh, C., Murdoch, W. J., and Yu, B. (2019). Hierarchical interpretations for neural network predictions. In *7th International Conference on Learning Representations (ICLR 2019)*.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES 2020)*, pages 180–186. ACM.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9259–9268. PMLR.

Sundararajan, M. and Najmi, A. (2020). The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR.

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

Tsang, M., Cheng, D., Liu, H., Feng, X., Zhou, E., and Liu, Y. (2020). Feature Interaction Interpretability: A Case for Explaining Ad-Recommendation Systems via Neural Interaction Detection. In *8th International Conference on Learning Representations (ICLR 2020)*.

Tsang, M., Cheng, D., and Liu, Y. (2018). Detecting Statistical Interactions from Neural Network Weights. In *6th International Conference on Learning Representations (ICLR 2018)*.

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020)*, pages 38–45. Association for Computational Linguistics.

Wright, M. N., Ziegler, A., and König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, 17:145.

**Organization of the Appendix.** Within the Appendix, we provide not only proofs for our theoretical analysis in Appendix E and further empirical results in Appendix G, but also give a table of frequently used symbols throughout the paper inAppendix A, provide Shapley-based interaction measures and other indices falling under the notion of cardinal interaction indices explicitly in Appendix B, provide further and more detailed pseudocode of our algorithmic approach SVARM-IQ in Appendix C, showcase our method at the popular special case of the Shapley Interaction index for pairs Appendix D, and describe the used models, datasets, and explanation tasks within our experimental setup in Appendix F. Appendix H contains the hardware details.

# A LIST OF SYMBOLS

| | |
|---|---|
| **Problem setting** | |
| $\mathcal{N}$ | Set of players |
| $\mathcal{N}_k$ | Set of all subsets of the player set with cardinality $k$ |
| $n$ | Number of players |
| $\nu$ | Value function |
| $B$ | Budget, number of allowed evaluations of $\nu$ |
| $k$ | Considered interaction order |
| $K$ | Interaction set |
| $\Delta_K(S)$ | Discrete derivative of players $K$ at coalition $S$ |
| $I_K$ | Cardinal interaction index of players $K$ |
| $\hat{I}_K$ | Estimated cardinal interaction index of players $K$ |
| $\mathcal{L}_k$ | Set of all coalition sizes at which interactions of order $k$ can occur |
| $\lambda_{k,\ell}$ | Weight of each coalition of size $\ell$ for interaction order $k$ |
| **SVARM-IQ** | |
| $I^W_{K,\ell}$ | Average worth of coalitions $S \cup W$ with $S \subseteq \mathcal{N} \setminus K$, $\lvert S\rvert = \ell$ and $W \subseteq K$ |
| $\hat{I}^W_{K,\ell}$ | Estimate of $I^W_{K,\ell}$ |
| $c^W_{K,\ell}$ | Number of samples observed for stratum $I^W_{K,\ell}$ |
| $\mathcal{S}_{\text{exp}}$ | Sizes for which all coalitions are evaluated for explicit stratum computation |
| $\mathcal{S}_{\text{imp}}$ | Sizes for which coalitions are sampled for implicit stratum estimation |
| $\mathcal{I}_{\text{exp}}$ | Set of all explicitly computed strata |
| $\mathcal{I}_{\text{imp}}$ | Set of all implicitly extimated strata |
| $\mathcal{L}^w_k$ | Set of implicit coalitions sizes $\ell$ depending on $W$ of $I^W_{K,\ell}$ |
| $P_k$ | Probability distribution over sizes $\{2, \ldots, n-2\}$ |
| $\bar{P}_k$ | Altered probability distribution over sizes $\{s_{\text{exp}}+1, \ldots, n-s_{\text{exp}}-1\}$ |
| $\bar{B}$ | Budget left for the sampling loop after `ComputeBorders` |
| $\tilde{B}$ | Budget left for the sampling loop after `ComputeBorders` and `WarmUp` |
| $\sigma^2_{K,W,\ell}$ | Variance of coalition worths in stratum $I^W_{K,\ell}$ |
| $r_{K,\ell,W}$ | Range of coalition worths in stratum $I^W_{K,\ell}$ |
| $\bar{m}^W_{K,\ell}$ | Number of samples with which $\hat{I}^W_{K,\ell}$ is updated after the warm-up |
| $m^W_{K,\ell}$ | Total number of samples with which $\hat{I}^W_{K,\ell}$ is updated |
| $A^W_{K,\ell,m}$ | $m$-th coalition used to update $\hat{I}^W_{K,\ell}$ |

Table 2: List of symbols used frequently throughout the paper.

# B CARDINAL INTERACTION INDICES AND THEIR WEIGHTS

All Shapley-based interaction indices and a few other game-theoretic measures of interaction can be captured under the notion of cardinal interaction indices (CII). We have stated this in Section 2 without presenting the aforementioned indices explicitly. We catch up on this by providing the weights $(\lambda_{k,\ell})_{\ell \in \mathcal{L}_k}$ of each index that is contained within the CII

$$I_K = \sum_{S \subseteq \mathcal{N} \setminus K} \lambda_{k,|S|} \cdot \Delta_K(S)$$

with discrete derivative

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

- Shapley Interaction index (SII) (Grabisch and Roubens, 1999):

$$\lambda_{k,\ell}^{\text{SII}} = \frac{1}{(n-k+1)\binom{n-k}{\ell}}$$

- Shapley-Taylor Interaction index (STI) (Sundararajan et al., 2020):

$$\lambda_{k,\ell}^{\text{STI}} = \frac{k}{n\binom{n-1}{\ell}}$$

- Faithful-Shapley Interaction index (FSI) (Tsai et al., 2023):

$$\lambda_{k,\ell}^{\text{FSI}} = \frac{(2k-1)!}{((k-1)!)^2} \cdot \frac{(n-\ell-1)!(\ell+k-1)!}{(n+k-1)!}$$

- Banzhaf Interaction index (BII) (Grabisch and Roubens, 1999):

$$\lambda_{k,\ell}^{\text{BII}} = \frac{1}{2^{n-k}}$$

For $k = 1$, the SII, STI, and FSI are identical and equal to the Shapley value:

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

which is why these are also called Shapley-based interactions. For a comprehensive overview of the axiomatic background justifying these indices, we refer to (Tsai et al., 2023) and (Fumagalli et al., 2023).

**n-SII Values.** The n-Shapley Values (n-SII) $\Phi^n$ were introduced by Bordt and von Luxburg (2023) as an extension of the Shapley interactions Lundberg et al. (2020) to higher orders. The n-SII constructs an interaction index for interactions up to size $n$, which is efficient, i.e. the sum of all interactions equals the full model $\nu(\mathcal{N})$. The n-SII are based on SII, $I^{\text{SII}}$, and aggregate SII up to order n. The highest interaction order of n-SII is always equal to SII. For every lower order, the n-SII values are constructed recursively, as

$$\Phi_K^n := \begin{cases} I_K^{\text{SII}}(S) & \text{if } |K| = n \\ \Phi_K^{n-1} + B_{n-|K|} \sum_{\substack{\tilde{K} \subseteq \mathcal{N} \setminus K \\ |K|+|\tilde{K}|=n}} I_{K \cup \tilde{K}}^{\text{SII}} & \text{if } |K| < n, \end{cases}$$

where the initial values are the SV $\Phi^1 \equiv \phi$ and $B_n$ are the Bernoulli numbers. It was shown Bordt and von Luxburg (2023) that n-SII yield an efficient index, i.e.

$$\sum_{\substack{K \subseteq \mathcal{N} \\ |K| \leq n}} \Phi_K^n = \nu(\mathcal{N}).$$

# C ADDITIONAL PSEUDOCODE

## C.1 Computing Border Sizes

We have only sketched the ComputeBorders procedure and will provide it now in full detail (see Algorithm 2). Its purpose is to determine the coalition sizes $\mathcal{S}_{\text{exp}}$ for which all coalitions are to be evaluated such that the corresponding strata are computed explicitly. We construct this set symmetrically, in the sense that a size $s_{\text{exp}}$ is determined such that all $\mathcal{S}_{\text{exp}} = \{0, \ldots, s_{\text{exp}}, n - s_{\text{exp}}, \ldots, n\}$, in other words: the smallest and the largest $s_{\text{exp}}$ many set sizes are included. Hence, we assume for simplicity that the initial probability distribution over sizes $P_k$ is symmetric, i.e., $P_k(s) = P_k(n-s)$, although it does not pose a challenge to extend this to any $P_k$ of arbitrary shape.

We start with $s_{\text{exp}} = 1$ and adjust the remaining budget $\bar{B}$ and the altered probability distribution over sizes $\bar{P}_k$. For each size $s$ being included into $\mathcal{S}_{\text{exp}}$, we set its probability mass to zero and upscale the remaining entries, effectively transferring probability mass from the border sizes to the middle. According to this procedure, ComputeBorders constructs $\bar{P}_k$ with

$$\bar{P}_k(s) = \frac{P_k(s)}{\sum_{s' \in \mathcal{S}_{\text{imp}}} P_k(s')} \text{ for all } s \in \mathcal{S}_{\text{imp}} \quad \text{and} \quad \bar{P}_k(s) = 0 \text{ for all } s \in \mathcal{S}_{\text{exp}}.$$

**Algorithm 2** ComputeBorders

$s_{\text{exp}} \leftarrow 1$
$\bar{B} \leftarrow B - 2n - 2$
$\bar{P}_k(0), \bar{P}_k(1), \bar{P}_k(n-1), \bar{P}_k(n) \leftarrow 0$
$\bar{P}_k(s) \leftarrow \frac{P_k(s)}{1 - P_k(0) - P_k(1) - P_k(n-1) - P_k(n)}$ for all $s \in \{2, \ldots, n-2\}$
**while** $s_{\text{exp}} + 1 \leq \frac{n}{2}$ **and** $\binom{n}{s_{\text{exp}}+1} \leq \bar{P}_k(s_{\text{exp}}+1) \cdot \bar{B}$ **do**
  $s_{\text{exp}} \leftarrow s_{\text{exp}} + 1$
  **if** $s_{\text{exp}} = \frac{n}{2}$ **then**
    $\bar{B} \leftarrow \bar{B} - \binom{n}{s_{\text{exp}}}$
    $\bar{P}_k \leftarrow \text{Unif}(0, n)$
  **else if** $s_{\text{exp}} = \frac{n-1}{2}$ **then**
    $\bar{B} \leftarrow \bar{B} - 2\binom{n}{s_{\text{exp}}}$
    $\bar{P}_k \leftarrow \text{Unif}(0, n)$
  **else**
    $\bar{B} \leftarrow \bar{B} - 2\binom{n}{s_{\text{exp}}}$
    $\bar{P}_k(s_{\text{exp}}) \leftarrow 0$
    $\bar{P}_k(n - s_{\text{exp}}) \leftarrow 0$
    $\bar{P}_k(s) \leftarrow \frac{\bar{P}_k(s)}{1 - 2\bar{P}(s_{\text{exp}})}$ for all $s \in \{s_{\text{exp}}+1, \ldots, n - s_{\text{exp}} - 1\}$
  **end if**
**end while**
$\mathcal{S}_{\text{exp}} \leftarrow \{0, \ldots, s_{\text{exp}}, n - s_{\text{exp}}, \ldots, n\}$
$\mathcal{S}_{\text{imp}} \leftarrow \{s_{\text{exp}}+1, \ldots, n - s_{\text{exp}} - 1\}$
**for** $s \in \mathcal{S}_{\text{exp}}$ **do**
  **for** $A \in \mathcal{N}_s$ **do**
    $v \leftarrow \nu(A)$
    **for** $K \in \mathcal{N}_k$ **do**
      $W \leftarrow A \cap K$
      $\ell \leftarrow s - |W|$
      $\hat{I}^W_{K,\ell} \leftarrow \hat{I}^W_{K,\ell} + \frac{v}{\binom{n-k}{\ell}}$
    **end for**
  **end for**
**end for**
**Output:** $\mathcal{S}_{\text{exp}}, \mathcal{S}_{\text{imp}}$

ComputeBorders iterates over sizes in increasing manner, checking whether the reminaing budget $\bar{B}$ is large enough such that the number of coalitions of the next size $s_{\text{exp}}+1$ considered is covered by the expected number of drawn coalitions with that size. As long as this holds true, $s_{\text{exp}}$ is incremented and $\bar{B}$ as well as $\bar{P}_k$ are adjusted. Note that thus not only $s_{\text{exp}}+1$ is added to $\mathcal{S}_{\text{exp}}$ but also $n-s_{\text{exp}}-1$. We distinguish between different cases, depending on whether the incremented $s_{\text{exp}}$ has reached the middle of the range of coalition sizes. In case of even $n$ this is $\frac{n}{2}$, otherwise $\frac{n-1}{2}$. As soon as $s_{\text{exp}}$ reaches that number, $\bar{P}_k(s)$ becomes irrelevant because then all coalitions of all sizes are being evaluated, leaving no strata to be estimated. In this case we simply set $\bar{P}_k$ to the uniform distribution such that it is well-defined.

After the computation of $\mathcal{S}_{\text{exp}}$, we evaluate all coalitions with cardinality $s \in \mathcal{S}_{\text{exp}}$. For each such coalition $A$ we update the estimate $\hat{I}_{K,\ell}^W$ with $W = A \cap K$ and $\ell = s - |W|$ according to our update mechanism. Since each stratum contains only coalitions of the same size, this leads to exactly computed strata representing the average of the contained coalitions' worths.

### C.2 Warm-up

The WarmUp (see Algorithm 3) procedure guarantees that each stratum estimate $\hat{I}_{K,\ell}^W$ with $I_{K,\ell}^W \in \mathcal{I}_{\text{imp}}$ is initialized with the worth of one sampled coalition. This is a natural requirement to facilitate our theoretical analysis in Appendix E. We achieve this algorithmically by iterating over all combinations of $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$. Each such combination specifies a stratum that is implicitly to be estimated. WarmUp draws for each stratum a coalition $A$ uniformly at random from the set of all coalitions of size $\ell$ and not containing any player of $K$. The estimate $\hat{I}_{K,\ell}^W$ is then set to the evaluated worth $\nu(A \cup W)$ and the counter of observed samples is set to one. The spent budget is:

$$
\begin{aligned}
|\mathcal{I}_{\text{imp}}| &= \binom{n}{k} \cdot \sum_{w=0}^{k} \binom{k}{w} |\mathcal{L}_k^{|w|}| \\
&= \binom{n}{k} \cdot \sum_{w=0}^{k} \binom{k}{w} |\{\max\{0, s_{\text{exp}}+1-w\}, \ldots, \min\{n-k, n-s_{\text{exp}}-1-w\}\}| \\
&= \binom{n}{k} \cdot \sum_{w=0}^{k} \binom{k}{w} \left(n - \max\{k, s_{\text{exp}}+1+w\} - \max\{0, s_{\text{exp}}+1-w\} + 1\right).
\end{aligned}
$$

**Algorithm 3** WarmUp

**for** $K \subseteq \mathcal{N}_k$ **do**
  **for** $W \subseteq K$ **do**
    **for** $\ell \in \mathcal{L}_k^{|W|}$ **do**
      Draw $A$ from $\{S \subseteq \mathcal{N} \setminus K \mid |S| = \ell\}$ uniformly at random
      $\hat{I}_{K,\ell}^W \leftarrow \nu(A \cup W)$
      $c_{K,\ell}^W \leftarrow 1$
    **end for**
  **end for**
**end for**

### C.3 Updating Strata Mean Estimates

In order to update the mean estimates $\hat{I}_{K,\ell}^W$ of the estimated strata incrementally with a single pass, thus not requiring to iterate over all previous samples, we use UpdateMean (see Algorithm 4). Besides the old estimate and the newly observed coalition worth $v_b$, this requires the number of observations made so far given by $c_{K,\ell}^W$.

**Algorithm 4** UpdateMean

**Input:** $\hat{I}_{K,\ell}^W$, $c_{K,\ell}^W$, $v_b$
**Output:** $\frac{\hat{I}_{K,\ell}^W \cdot c_{K,\ell}^W + v_b}{c_{K,\ell}^W + 1}$

# D THE SPECIAL CASE OF PAIRWISE SHAPLEY-INTERACTIONS

We stated our approximation algorithm SVARM-IQ for all CII and any order $k$. Since the Shapley Interaction index (SII) for pairs, i.e., $k = 2$, is the most popular among them, we provide a description of SVARM-IQ and the pseudocode (see Algorithm 5) for that specific case, leading to a simpler presentation of our approach.

The SII of a pair of players $\{i, j\} \in \mathcal{N}_2$ is given by

$$I_{i,j}^{\text{SII}} = \sum_{S \subseteq \mathcal{N} \setminus \{i,j\}} \frac{1}{(n-1)\binom{n-2}{|S|}} \left[\nu(S \cup \{i,j\}) - \nu(S \cup \{i\}) - \nu(S \cup \{j\}) + \nu(S)\right].$$

Now, our approach stratifies the discrete derivatives $\Delta_{i,j}(S)$ by size and splits them into multiple strata, yielding the following representation of the SII:

$$I_{i,j}^{\text{SII}} = \frac{1}{n-1} \sum_{\ell=0}^{n-2} I_{i,j,\ell}^{\{i,j\}} - I_{i,j,\ell}^{\{i\}} - I_{i,j,\ell}^{\{i\}} + I_{i,j,\ell}^{\emptyset},$$

with strata terms for all $W \subseteq \{i, j\}$ and $\ell \in \mathcal{L}_2 := \{0, \ldots, n-2\}$:

$$I_{i,j,\ell}^{W} := \frac{1}{\binom{n-2}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \{i,j\} \\ |S| = \ell}} \nu(S \cup W).$$

We keep a stratum estimate $\hat{I}_{i,j,\ell}^{W}$ for each pair $i$ and $j$, size $\ell \in \mathcal{L}_2$, and subset $W \subseteq \{i, j\}$. Subsequently, the aggregation of the strata estimates, which we obtain during sampling, provides the desired SII estimate:

$$\hat{I}_{i,j}^{\text{SII}} := \frac{1}{n-1} \sum_{\ell=0}^{n-2} \hat{I}_{i,j,\ell}^{\{i,j\}} - \hat{I}_{i,j,\ell}^{\{i\}} - \hat{I}_{i,j,\ell}^{\{i\}} + \hat{I}_{i,j,\ell}^{\emptyset}.$$

For each sampled coalition $A$ of size $|A| = a$, the update mechanism needs to distinguish between only 4 cases. For each pair $i$ and $j$ it updates:

- $\hat{I}_{i,j,a-2}^{\{i,j\}}$ if $i, j \in A$,
- $\hat{I}_{i,j,a-1}^{\{i\}}$ if $i \in A$ but $j \notin A$,
- $\hat{I}_{i,j,a-1}^{\{j\}}$ if $j \in A$ but $i \notin A$, or
- $\hat{I}_{i,j,a}^{\emptyset}$ if $i, j \notin A$.

This case distinction is still captured by computing $W = A \cap K$, $\ell = a - |W|$, and updating $\hat{I}_{i,j,\ell}^{W}$.

**Algorithm 5** SVARM-IQ (for the Shapley Interaction index of order $k = 2$)

**Input:** $(\mathcal{N}, \nu)$, $B \in \mathbb{N}$
$\hat{I}^{\emptyset}_{i,j,\ell}, \hat{I}^{\{i\}}_{i,j,\ell}, \hat{I}^{\{j\}}_{i,j,\ell}, \hat{I}^{\{i,j\}}_{i,j,\ell} \leftarrow 0$ for all $\{i,j\} \in \mathcal{N}_2, \ell \in \mathcal{L}_2$
$c^{\emptyset}_{i,j,\ell}, c^{\{i\}}_{i,j,\ell}, c^{\{j\}}_{i,j,\ell}, c^{\{i,j\}}_{i,j,\ell} \leftarrow 0$ for all $\{i,j\} \in \mathcal{N}_2, \ell \in \mathcal{L}_2$
ComputeBorders
$\bar{B} \leftarrow B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s}$
**for** $b = 1, \ldots, \bar{B}$ **do**
  Draw size $a_b \in \mathcal{S}_{\text{imp}} \sim \bar{P}_k$
  Draw $A_b$ from $\{S \subseteq \mathcal{N} \mid |S| = a_b\}$ uniformly at random
  $v_b \leftarrow \nu(A_b)$
  **for** $\{i,j\} \in \mathcal{N}_2$ **do**
    $W \leftarrow A_b \cap \{i,j\}$
    $\ell \leftarrow a_b - |W|$
    $\hat{I}^{W}_{i,j,\ell} \leftarrow \frac{\hat{I}^{W}_{i,j,\ell} \cdot c^{W}_{i,j,\ell} + v_b}{c^{W}_{i,j,\ell} + 1}$
    $c^{W}_{i,j,\ell} \leftarrow c^{W}_{i,j,\ell} + 1$
  **end for**
**end for**
$\hat{I}_{i,j} \leftarrow \frac{1}{n-1} \sum_{\ell=0}^{n-2} \hat{I}^{\{i,j\}}_{i,j,\ell} - \hat{I}^{\{i\}}_{i,j,\ell} - \hat{I}^{\{j\}}_{i,j,\ell} + \hat{I}^{\emptyset}_{i,j,\ell}$ for all $\{i,j\} \in \mathcal{N}_2$
**Output:** $\hat{I}_{i,j}$ for all $\{i,j\} \in \mathcal{N}_2$

# E PROOFS

In the following, we give the proofs to our theoretical results in Section 4. We start by defining and revisiting some helpful notation and stating our assumptions.

**Notation:**

- Let $\mathcal{L}_k := \{0, \dots, n-k\}$.
- Let $\mathcal{L}_k^{|W|} := \{\ell \in \mathcal{L}_k \mid \ell + |W| \in \mathcal{S}_{\text{imp}}\} = [\max\{0, s_{\text{exp}} + 1 - w\}, \min\{n-k, n-s_{\text{exp}}-1\}]$ for any $W \subseteq K \in \mathcal{N}_k$.
- Let $\tilde{B} = B - \sum_{s \in \mathcal{S}_{\text{exp}}} \binom{n}{s} - |\mathcal{I}_{\text{imp}}|$ be the available budget left for the sampling loop after the completion of ComputeBorders and WarmUp.
- For all $K \in \mathcal{N}_k$ with $\ell \in \mathcal{L}_k$, let $A_{K,\ell}$ be a random set with $\mathbb{P}(A_{K,\ell} = S) = \frac{1}{\binom{n-k}{\ell}}$ for all $S \subseteq \mathcal{N} \setminus K$ with $|S| = \ell$.
- For all $K \in \mathcal{N}_k$ with $W \subseteq K$ and $\ell \in \mathcal{L}_k^w$:
  - Let $\sigma^2_{K,\ell,W} := \mathbb{V}[\nu(A_{K,\ell} \cup W)]$ be the strata variance.
  - Let $r_{K,\ell,W} := \max\limits_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W) - \min\limits_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S| = \ell}} \nu(S \cup W)$ be the strata range.
  - Let $\bar{m}^W_{K,\ell} := \#\{b \mid |A_b| = \ell + |W|, A_b \cap K = W\}$ be the number of samples with which $\hat{I}^W_{K,\ell}$ is updated during the sampling loop.
  - Let $m^W_{K,\ell} := \bar{m}^W_{K,\ell} + 1$ be the total number of samples with which $\hat{I}^W_{K,\ell}$ is updated.
  - Let $A^W_{K,\ell,m}$ be the $m$-th coalition used to update $I^W_{K,\ell}$.
- For all $K \in \mathcal{N}_k$ let $R_K := \sum\limits_{W \subseteq K} \sum\limits_{\ell \in \mathcal{L}_k^{|W|}} r_{K,\ell,W}$.
- Let $\gamma_k$ be $\gamma_2 := 2(n-1)^2$ for $k = 2$ and $\gamma_k := n^{k-1}(n-k+1)^2$ for $k \geq 3$.

**Assumptions:**

- $\tilde{B} > 0$
- $n \geq 4$
- $B < 2^n$

The lower bound on the leftover budget $\tilde{B}$ is necessary to ensure the completion of ComputeBorders and WarmUp, and that at least one coalition is sampled during the sampling loop. The assumption on $n$ arises from the fact that ComputeBorders automatically evaluates the worth of all coalitions having size $0, 1, n-1$ or $n$. Hence, all CII values are computed exactly for $n = 3$. Our considered problem statement becomes trivial for $n \leq 2$. Likewise, in order to avoid triviality, we demand the budget to be lower than the total number of coalitions $2^n$. Otherwise, all CII values will be computed exactly by ComputeBorders and the approximation problem vanishes. This allows us to state $\mathcal{S}_{\text{imp}} \neq \emptyset$ and $\mathcal{I}_{\text{imp}} \neq \emptyset$.

### E.1 Unbiasedness

**Lemma E.1.** *All strata estimates $\hat{I}^W_{K,\ell}$ are unbiased, i.e., for all $K \in \mathcal{N}_k$, $W \subseteq K$, $\ell \in \mathcal{L}_k$:*

$$\mathbb{E}\left[\hat{I}^W_{K,\ell}\right] = I^W_{K,\ell}.$$

*Proof.* The statement trivially holds for all strata explicitly computed by ComputeBorders. Thus, we consider the remaining strata which are estimated via sampling. Fix any $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$. Due to the uniform sampling of eligible coalitions once the size is fixed, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\hat{I}_{K,\ell}^W \mid m_{K,\ell}^W\right] \\
&= \frac{1}{m_{K,\ell}^W} \sum_{m=1}^{m_{K,\ell}^W} \mathbb{E}\left[\nu(A_{K,\ell,m}^W) \mid m_{K,\ell}^W\right] \\
&= \frac{1}{m_{K,\ell}^W} \sum_{m=1}^{m_{K,\ell}^W} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S|=\ell}} \mathbb{P}(A_{K,\ell,m}^W = S \cup W \mid |A_{K,\ell,m}^W| = \ell + |W|, A_{K,\ell,m}^W \cap K = W) \cdot \nu(S \cup W) \\
&= \frac{1}{m_{K,\ell}^W} \sum_{m=1}^{m_{K,\ell}^W} \sum_{\substack{S \subseteq \mathcal{N} \setminus K \\ |S|=\ell}} \frac{1}{\binom{n-k}{\ell}} \cdot \nu(S \cup W) \\
&= \frac{1}{m_{K,\ell}^W} \sum_{m=1}^{m_{K,\ell}^W} I_{K,\ell}^W \\
&= I_{K,\ell}^W.
\end{aligned}
$$

Note that the set $A_{K,\ell,m}^W$ has cardinality $\ell + |W|$ and fulfills $A_{K,\ell,m}^W \cap K = W$ by definition. Otherwise, it would not be used to update $\hat{I}_{K,\ell}^W$. Since `WarmUp` gathers one sample for each stratum estimate, it guarantees $m_{K,\ell}^W \geq 1$. Thus the above terms are well defined. Finally, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[\hat{I}_{K,\ell}^W\right] &= \sum_{m=1}^{\bar{B}+1} \mathbb{E}\left[\hat{I}_{K,\ell}^W \mid m_{K,\ell}^W = m\right] \cdot \mathbb{P}(m_{K,\ell}^W = m) \\
&= \sum_{m=1}^{\bar{B}+1} I_{K,\ell}^W \cdot \mathbb{P}(m_{K,\ell}^W = m) \\
&= I_{K,\ell}^W.
\end{aligned}
$$

□

**Theorem 4.1.** *The CII estimates returned by SVARM-IQ are unbiased for all $K \in \mathcal{N}_k$, i.e.,*

$$
\mathbb{E}\left[\hat{I}_K\right] = I_K.
$$

*Proof.* We have already proven the unbiasedness of all strata estimates with Lemma E.1. Thus, we obtain for all $K \in \mathcal{N}_k$:

$$
\begin{aligned}
\mathbb{E}\left[\hat{I}_K\right] &= \mathbb{E}\left[\sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \hat{I}_{K,\ell}^W\right] \\
&= \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot \mathbb{E}\left[\hat{I}_{K,\ell}^W\right] \\
&= \sum_{\ell=0}^{n-k} \binom{n-k}{\ell} \lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot I_{K,\ell}^W \\
&= I_K.
\end{aligned}
$$

□

### E.2 Sample Numbers

Form now on, we distinguish between the special case of order $k = 2$ and all others $k \geq 3$, allowing us to give tighter bounds for the former. Hence, we introduce $\gamma_k$ for all $k \geq 2$ with

$$\gamma_2 = 2(n-1)^2 \text{ and } \gamma_k = n^{k-1}(n-k+1)^2 \text{ for all } k \geq 3.$$

**Lemma E.2.** *The number of samples $\bar{m}^W_{K,\ell}$ collected for the strata estimate $\hat{I}^W_{K,\ell}$ of any fixed player set $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$ collected during the sampling loop is binomially distributed with an expected value of at least*

$$\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] \geq \frac{\tilde{B}}{\gamma_k}.$$

*Proof.* The number of collected samples during the sample loop, i.e. $\bar{m}^W_{K,\ell}$, is binomially distributed because in each iteration the stratum $\hat{I}^W_{K,\ell}$ has the same probability to be updated and the sampled coalitions are independent of each other across the iterations. The number of iterations is $\tilde{B}$ and the condition for an update of $\hat{I}^W_{K,\ell}$ is that the sampled set $A_b$ fulfills $|A_b| = a_b = \ell + |W|$ and $A_b \cap K = W$. This happens with a probability of:

$$\begin{aligned}
&\mathbb{P}(a_b = \ell + |W|, A_b \cap K = W) \\
&= \mathbb{P}(A_b \cap K = W \mid a_b = \ell + |W|) \cdot \mathbb{P}(a_b = \ell + |W|) \\
&= \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot \bar{P}_k(\ell + |W|).
\end{aligned}$$

Hence, we obtain $\bar{m}^W_{K,\ell} \sim Bin\left(\tilde{B}, \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot \bar{P}_k(\ell + |W|)\right)$. This yields

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] &= \tilde{B} \cdot \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot \bar{P}_k(\ell + |W|) \\
&\geq \tilde{B} \cdot \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_k(\ell + |W|).
\end{aligned}$$

Note that $\bar{P}_k(\ell+|W|) \geq P_k(\ell+|W|)$ holds true for all $\ell$ and $W \subseteq K$ with $\ell+|W| \in \mathcal{S}_{\text{imp}}$ because for these sizes, from which coalitions are sampled, $\bar{P}_k$ can only gain probability mass in comparison to $P_k$. More precisely, for all $s \in \mathcal{S}_{\text{imp}}$ we have

$$\bar{P}_k(s) = \frac{P_k(s)}{\sum_{s' \in \mathcal{S}_{\text{imp}}} P_k(s')} \geq \frac{P_k(s)}{\sum_{s' \in \mathcal{S}_{\text{exp}}} P_k(s') + \sum_{s' \in \mathcal{S}_{\text{imp}}} P_k(s')} = P_k(s).$$

We continue to prove our statement for the case of $k = 2$ and any fixed $K = \{i, j\}$ by giving a lower bound for the expected value of $\bar{m}^W_{i,j,\ell}$. Inserting $k = 2$, we can further write

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{i,j,\ell}\right] &\geq \tilde{B} \cdot \frac{\binom{n-2}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_2(\ell + |W|) \\
&= \frac{\tilde{B}}{n(n-1)} \cdot \frac{(\ell+|W|)!}{\ell!} \cdot \frac{(n-\ell-|W|)!}{(n-\ell-2)!} \cdot P_2(\ell + |W|).
\end{aligned}$$

Let

$$f(\ell, w) := \frac{(\ell+w)!}{\ell!} \cdot \frac{(n-\ell-w)!}{(n-\ell-2)!} = \begin{cases} (n-\ell)(n-\ell-1) & \text{if } w = 0 \\ (\ell+1)(n-\ell-1) & \text{if } w = 1 \\ (\ell+1)(\ell+2) & \text{if } w = 2 \end{cases}.$$

In the following, we derive the lower bound $f(\ell, |W|) \cdot P_2(\ell + |W|) \geq \frac{n}{2(n-1)}$ for all $|W| \in \{0, 1, 2\}$ and $\ell \in \mathcal{L}_2^{|W|}$ by distinguishing over different cases of $n$, $\ell$, and $|W|$ and exploiting our tailored distribution $P_2$.

For odd $n$, $\ell + |W| \leq \frac{n-1}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{\ell(\ell-1)} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)^2}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For odd $n$, $\ell + |W| \leq \frac{n-1}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(\ell+1)\ell} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)(n+1)}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For odd $n$, $\ell + |W| \leq \frac{n-1}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(\ell+2)(\ell+1)} \cdot \frac{n-1}{2(n-3)} = \frac{n-1}{2(n-3)} \geq \frac{n}{2(n-1)}$$

For odd $n$, $\ell + |W| \geq \frac{n+1}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{(n-\ell)(n-\ell-1)} \cdot \frac{n-1}{2(n-3)} = \frac{n-1}{2(n-3)} \geq \frac{n}{2(n-1)}$$

For odd $n$, $\ell + |W| \geq \frac{n+1}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(n-\ell-1)(n-\ell-2)} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)(n+1)}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For odd $n$, $\ell + |W| \geq \frac{n+1}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(n-\ell-2)(n-\ell-3)} \cdot \frac{n-1}{2(n-3)} \geq \frac{(n-1)(n+1)}{2(n-3)^2} \geq \frac{n}{2(n-1)}$$

For even $n$, $\ell + |W| \leq \frac{n-2}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{\ell(\ell-1)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n^2(n+2)}{2(n-4)(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even $n$, $\ell + |W| \leq \frac{n-2}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(\ell+1)\ell} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n(n-2)(n+2)}{2(n-4)(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even $n$, $\ell + |W| \leq \frac{n-2}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(\ell+2)(\ell+1)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} = \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even $n$, $\ell + |W| \geq \frac{n}{2}$, and $|W| = 0$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(n-\ell)(n-\ell-1)}{(n-\ell)(n-\ell-1)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} = \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even $n$, $\ell + |W| \geq \frac{n}{2}$, and $|W| = 1$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(n-\ell-1)}{(n-\ell-1)(n-\ell-2)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n^2}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

For even $n$, $\ell + |W| \geq \frac{n}{2}$, and $|W| = 2$:

$$f(\ell, |W|) \cdot P_2(\ell + |W|) = \frac{(\ell+1)(\ell+2)}{(n-\ell-2)(n-\ell-3)} \cdot \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n^2-2n}{2(n^2-4n+2)} \geq \frac{n}{2(n-1)}$$

This allows us to conclude:

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{i,j,\ell}\right] &\geq \tilde{B} \cdot \frac{\binom{n-2}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_2(\ell + |W|) \\
&= \frac{\tilde{B}}{n(n-1)} \cdot \frac{(\ell+|W|)!}{\ell!} \cdot \frac{(n-\ell-|W|)!}{(n-\ell-2)!} \cdot P_2(\ell + |W|) \\
&\geq \frac{\tilde{B}}{n(n-1)} \cdot \frac{n}{2(n-1)} \\
&= \frac{\tilde{B}}{\gamma_2}.
\end{aligned}$$

Next, we turn our attention to the case of $k \geq 3$. Inserting the uniform distribution for $P_k$, we can write for the expected number of samples:

$$\begin{aligned}
\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] &\geq \tilde{B} \cdot \frac{\binom{n-k}{\ell}}{\binom{n}{\ell+|W|}} \cdot P_k(\ell + |W|) \\
&= \tilde{B} \cdot \frac{(n-k)!}{n!} \cdot \frac{(\ell+|W|)!}{\ell!} \cdot \frac{(n-\ell-|W|)!}{(n-\ell-k)!} \cdot \frac{1}{n-3} \\
&\geq \tilde{B} \cdot \frac{(n-k)!}{n!} \cdot \frac{1}{n-3}.
\end{aligned}$$

In the following we prove that $\frac{(n-k)!}{n!} \cdot \frac{1}{n-3} \geq \frac{1}{n^{k-1}(n-k+1)^2}$. First, we obtain the equivalent inequality

$$n^{k-1}(n-k+1) \geq (n-3) \prod_{i=n-k+2}^{n} i.$$

Note that we have $n \geq k$ at all times. The inequality obviously holds true for all $k \leq 4$. We prove its correctness for $k \geq 5$ by induction over $k$. We start with the induction base at $k = 5$:

$$\begin{aligned}
& n^{k-1}(n-k+1) \geq (n-3) \prod_{i=n-k+2}^{n} i \\
\Leftrightarrow\ & n^3(n-4) \geq (n-1)(n-2)(n-3)^2 \\
\Leftrightarrow\ & 5n^3 + 39n \geq 29n^2 + 18.
\end{aligned}$$

The resulting equality is obviously fulfilled by all $n \geq 5$. Next, we conduct the induction step by considering the inequality for $k+1$ with $k \geq 5$:

$$\begin{aligned}
& n^k(n-k) \\
&= \frac{n(n-k)}{n-k+1} \cdot n^{k-1}(n-k+1) \\
&\geq \frac{n(n-k)}{n-k+1} \cdot (n-3) \prod_{i=n-k+2}^{n} i \\
&\geq (n-k+1) \cdot (n-3) \prod_{i=n-k+2}^{n} i \\
&= (n-3) \prod_{i=n-k+1}^{n} i.
\end{aligned}$$

With the inequality proven, we finally obtain the desired lower bound for the expectation of $\bar{m}^W_{K,\ell}$:

$$\begin{aligned}\mathbb{E}\left[\bar{m}^W_{K,\ell}\right] &\geq \tilde{B} \cdot \frac{(n-k)!}{n!} \cdot \frac{1}{n-3} \\ &\geq \frac{\tilde{B}}{n^{k-1}(n-k+1)^2} \\ &= \frac{\tilde{B}}{\gamma_k}.\end{aligned}$$

□

**Lemma E.3.** *The expected inverted total sample number of the strata estimate $\hat{I}^W_{K,\ell}$ for any fixed $K \in \mathcal{N}_k$, $W \subseteq K$, and $\ell \in \mathcal{L}_k^{|W|}$ is bounded by*

$$\mathbb{E}\left[\frac{1}{m^W_{K,\ell}}\right] \leq \frac{\gamma_k}{\tilde{B}}.$$

*Proof.* In the following, we apply equation (3.4) in (Chao and Strawderman, 1972), stating

$$\mathbb{E}\left[\frac{1}{X+1}\right] = \frac{1-(1-p)^{m+1}}{(m+1)p} \leq \frac{1}{mp} = \frac{1}{\mathbb{E}[X]},$$

for any binomially distributed random variable $X \sim Bin(m,p)$. Due to `WarmUp` we have $m^W_{K,\ell} = \bar{m}^W_{K,\ell} + 1$, since it guarantees exactly one sample for each stratum. Next, Lemma E.2 allows us to substitute $X$ with $\bar{m}^W_{K,\ell}$ and we obtain:

$$\mathbb{E}\left[\frac{1}{m^W_{K,\ell}}\right] = \left[\frac{1}{\bar{m}^W_{K,\ell}+1}\right] \leq \frac{1}{\mathbb{E}\left[\bar{m}^W_{K,\ell}\right]} \leq \frac{\gamma_k}{\tilde{B}}.$$

□

### E.3 Variance and Mean Squared Error

**Lemma E.4.** *For any $K \in \mathcal{N}_k$, given the sample numbers $m^W_{K,\ell}$ for all $W \subseteq K$ and $\ell \in \mathcal{L}_k^{|W|}$, the variance of the estimate $\hat{I}_K$ is given by*

$$\mathbb{V}\left[\hat{I}_K \mid \left(m^W_{K,\ell}\right)_{\ell \in \mathcal{L}, W \subseteq K}\right] = \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \cdot \frac{\sigma^2_{K,\ell,W}}{m^W_{K,\ell}}.$$

*Proof.* First, we split the variance of $\hat{I}_K$ with the help of Bienaymé's identity into the variances of the strata estimates and their covariances. Then we make use of the fact that each sample to update a stratum is effectively drawn uniformly:

$$
\begin{aligned}
&\mathbb{V}\left[\hat{I}_K \mid \left(m^W_{K,\ell}\right)_{\ell\in\mathcal{L},W\subseteq K}\right] \\
&= \mathbb{V}\left[\sum_{\ell=0}^{n-k}\binom{n-k}{\ell}\lambda_{k,\ell}\sum_{W\subseteq K}(-1)^{k-|W|}\cdot\hat{I}^W_{K,\ell} \mid \left(m^W_{K,\ell}\right)_{\ell\in\mathcal{L},W\subseteq K}\right] \\
&= \mathbb{V}\left[\sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}\lambda_{k,\ell}\cdot(-1)^{k-|W|}\cdot\hat{I}^W_{K,\ell} \mid \left(m^W_{K,\ell}\right)_{\ell\in\mathcal{L},W\subseteq K}\right] \\
&= \sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\mathbb{V}\left[\hat{I}^W_{K,\ell} \mid m^W_{K,\ell}\right] \\
&\quad + \sum_{\substack{\ell\in\mathcal{L}_k\\ W\subseteq K}}\sum_{\substack{\ell'\in\mathcal{L}_k\\ W'\subseteq K\\ \ell\neq\ell'\vee W\neq W'}}\binom{n-k}{\ell}\binom{n-k}{\ell'}\lambda_{k,\ell}\lambda_{k,\ell'}\cdot(-1)^{2k-|W|-|W'|}\cdot\mathrm{Cov}\left(\hat{I}^W_{K,\ell},\hat{I}^{W'}_{K,\ell'} \mid m^W_{K,\ell},m^{W'}_{K,\ell'}\right) \\
&= \sum_{\ell=0}^{n-k}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\sum_{W\subseteq K}\mathbb{V}\left[\hat{I}^W_{K,\ell} \mid m^W_{K,\ell}\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\mathbb{V}\left[\hat{I}^W_{K,\ell} \mid m^W_{K,\ell}\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\mathbb{V}\left[\frac{1}{m^W_{K,\ell}}\sum_{m=1}^{m^W_{K,\ell}}\nu(A^W_{K,\ell,m}) \mid m^W_{K,\ell}\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\cdot\frac{\sigma^2_{K,\ell,W}}{m^W_{K,\ell}}.
\end{aligned}
$$

The strata estimates $\hat{I}^W_{K,\ell}$ and $\hat{I}^{W'}_{K,\ell'}$ are independent for $W \neq W'$ or $\ell = \ell'$ because each sampled coalition $A_b$ can only be used to update one estimate. Consequently, their covariance is zero. Finally, the variances of the estimates for the explicitly calculated strata are zero and thus eliminated. □

**Theorem 4.2.** *For any $K \in \mathcal{N}_k$ the variance of the estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by*

$$
\mathbb{V}\left[\hat{I}_K\right] \leq \frac{\gamma_k}{\tilde{B}}\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\sigma^2_{K,\ell,W}.
$$

*Proof.* We combine the variance of each estimate variance conditioned on the sample numbers given by Lemma E.4 with the bound on the expected inverted total sample numbers given by Lemma E.3:

$$
\begin{aligned}
\mathbb{V}\left[\hat{I}_K\right] &= \mathbb{E}_{\left(m^W_{K,\ell}\right)_{\ell\in\mathcal{L},W\subseteq K}}\left[\mathbb{V}\left[\hat{I}_K \mid \left(m^W_{K,\ell}\right)_{\ell\in\mathcal{L},W\subseteq K}\right]\right] \\
&= \mathbb{E}_{\left(m^W_{K,\ell}\right)_{\ell\in\mathcal{L},W\subseteq K}}\left[\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\cdot\frac{\sigma^2_{K,\ell,W}}{m^W_{K,\ell}}\right] \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\sigma^2_{K,\ell,W}\cdot\mathbb{E}\left[\frac{1}{m^W_{K,\ell}}\right] \\
&\leq \frac{\gamma_k}{\tilde{B}}\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}^2\lambda^2_{k,\ell}\sigma^2_{K,\ell,W}.
\end{aligned}
$$

□

**Corollary 4.3.** *For any $K \in \mathcal{N}_k$ the mean squared error of the estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by*

$$\mathbb{E}\left[\left(\hat{I}_K - I_K\right)^2\right] \leq \frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.$$

*Proof.* The bias-variance decomposition allows us to decompose the mean squared error into the bias of $\hat{I}_K$ and its variance. Since we have shown the estimate's unbiasedness in Theorem 4.1, we can reduce it to its variance bounded in Theorem 4.2:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{I}_K - I_K\right)^2\right] &= \left(\mathbb{E}\left[\hat{I}_K - I_K\right]\right)^2 + \mathbb{V}\left[\hat{I}_K\right] \\
&= \mathbb{V}\left[\hat{I}_K\right] \\
&\leq \frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.
\end{aligned}$$

□

### E.4 Threshold Exceedence Probability

**Corollary 4.4.** *For any $K \in \mathcal{N}_k$ and fixed $\varepsilon > 0$ the absolute error of the estimate $\hat{I}_K$ returned by SVARM-IQ exceeds $\varepsilon$ with a probability of at most*

$$\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \leq \frac{\gamma_k}{\varepsilon^2 \tilde{B}} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.$$

*Proof.* We apply Chebychev's inequality and make use of the variance bound in Theorem 4.2:

$$\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \leq \frac{\mathbb{V}\left[\hat{I}_k\right]}{\varepsilon^2} \leq \frac{\gamma_k}{\varepsilon^2 \bar{B}_k} \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \binom{n-k}{\ell}^2 \lambda_{k,\ell}^2 \sigma_{K,\ell,W}^2.$$

□

**Lemma E.5.** *For the stratum estimate $\hat{I}_{K,\ell}^W$ of any $K \in \mathcal{N}_k$ with $W \subseteq K$, $\ell \in \mathcal{L}_k^{|W|}$, and some fixed $\varepsilon > 0$ holds*

$$\mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon \mid m_{K,\ell}^W\right) \leq 2\exp\left(-\frac{2m_{K,\ell}^W \varepsilon^2}{r_{K,\ell,W}^2}\right).$$

*Proof.* We combine Hoeffding's inequality with the unbiasedness of the strata estimates shown in Lemma E.1 and obtain:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon \mid m_{K,\ell}^W\right) \\
&= \mathbb{P}\left(|\hat{I}_{K,\ell}^W - \mathbb{E}[\hat{I}_{K,\ell}^W]| \geq \varepsilon \mid m_{K,\ell}^W\right) \\
&= \mathbb{P}\left(\left|\sum_{m=1}^{m_{K,\ell}^W} \nu(A_{S,\ell,m}^W) - \mathbb{E}\left[\sum_{m=1}^{m_{K,\ell}^W} \nu(A_{K,\ell,m}^W)\right]\right| \geq m_{K,\ell}^W \varepsilon \mid m_{K,\ell}^W\right) \\
&\leq 2\exp\left(-\frac{2m_{K,\ell}^W \varepsilon^2}{r_{K,\ell,W}^2}\right).
\end{aligned}$$

□

**Lemma E.6.** *For any $K \in \mathcal{N}_k$ with $W \subseteq K$, $\ell \in \mathcal{L}_k^{|W|}$ and some fixed $\varepsilon > 0$ holds*

$$\mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon\right) \leq \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right) - 1}.$$

*Proof.* We start by deriving with Hoeffding's inequality and Lemma E.2 a bound on the probability that $\bar{m}_{K,\ell}^W$ falls below $\frac{\tilde{B}}{2\gamma_k}$:

$$\begin{aligned}
&\mathbb{P}\left(\bar{m}_{K,\ell}^W \leq \frac{\tilde{B}}{2\gamma_k}\right) \\
&= \mathbb{P}\left(\mathbb{E}\left[\bar{m}_{K,\ell}^W\right] - \bar{m}_{K,\ell}^W \geq \mathbb{E}\left[\bar{m}_{K,\ell}^W\right] - \frac{\tilde{B}}{2\gamma_k}\right) \\
&\leq \exp\left(-\frac{2\left(\mathbb{E}\left[\bar{m}_{K,\ell}^W\right] - \frac{\tilde{B}}{2\gamma_k}\right)^2}{\tilde{B}}\right) \\
&\leq \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right).
\end{aligned}$$

Further, we show with Lemma E.5 another statement:

$$\begin{aligned}
&\sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\tilde{B}+1} \mathbb{P}\left(|\hat{I}_{K,\ell}^W - I_{K,\ell}^W| \geq \varepsilon \mid m_{K,\ell}^W = m\right) \\
&\leq 2 \sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\tilde{B}+1} \exp\left(-\frac{2m\varepsilon^2}{r_{K,\ell,W}^2}\right) \\
&= 2 \sum_{m=0}^{\tilde{B}+1} \exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^m - 2 \sum_{m=0}^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor} \exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^m \\
&= 2\frac{\exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor} - \exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\tilde{B}+1}}{\exp\left(\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right) - 1} \\
&\leq 2\frac{\exp\left(-\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{r_{K,\ell,W}^2}\right) - 1}.
\end{aligned}$$

Finally, we combine both intermediate results and obtain:

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon\right) \\
&\leq \sum_{m=1}^{\tilde{B}+1} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \cdot \mathbb{P}\left(m^W_{K,\ell} = m\right) \\
&= \sum_{m=1}^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \cdot \mathbb{P}\left(m^W_{K,\ell} = m\right) \\
&\quad + \sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\bar{B}_k+1} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{K,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \cdot \mathbb{P}\left(m^W_{K,\ell} = m\right) \\
&\leq \mathbb{P}\left(m^W_{K,\ell} \leq \left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor\right) + \sum_{m=\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor + 1}^{\tilde{B}+1} \mathbb{P}\left(|\hat{I}^W_{K,\ell} - I^W_{S,\ell}| \geq \varepsilon \mid m^W_{K,\ell} = m\right) \\
&\leq \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{r^2_{K,\ell,W}}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{r^2_{K,\ell,W}}\right) - 1}.
\end{aligned}
$$

□

**Theorem 4.5.** *For any $K \in \mathcal{N}_k$ and fixed $\varepsilon > 0$ the absolute error of the estimate $\hat{I}_K$ exceeds $\varepsilon$ with probability of at most*

$$
\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \leq \sum_{W \subseteq K} \sum_{\ell \in \mathcal{L}_k^{|W|}} \exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2 \lambda^2_{k,\ell} R_K^2}\right)^{\left\lfloor \frac{\tilde{B}}{2\gamma_k} \right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2 \lambda^2_{k,\ell} R_K^2}\right) - 1}.
$$

*Proof.* We derive the result by applying Lemma E.6 and utilizing the fact that for all explicitly computed strata $I_{K,\ell}^W \in \mathcal{I}_{\text{exp}}$ holds $\hat{I}_{K,\ell}^W = I_{K,\ell}^W$:

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{I}_K - I_K| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\left|\sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}\lambda_{k,\ell}(-1)^{k-|W|}\left(\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right)\right| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(\sum_{\ell=0}^{n-k}\sum_{W\subseteq K}\binom{n-k}{\ell}\lambda_{k,\ell}\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\binom{n-k}{\ell}\lambda_{k,\ell}\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \varepsilon\right) \\
&\leq \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\mathbb{P}\left(\binom{n-k}{\ell}\lambda_{k,\ell}\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \frac{\varepsilon r_{K,\ell,W}}{R_K}\right) \\
&= \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\mathbb{P}\left(\left|\hat{I}_{K,\ell}^W - I_{K,\ell}^W\right| \geq \frac{\varepsilon r_{K,\ell,W}}{\binom{n-k}{\ell}\lambda_{k,\ell}R_K}\right) \\
&\leq \sum_{W\subseteq K}\sum_{\ell\in\mathcal{L}_k^{|W|}}\exp\left(-\frac{\tilde{B}}{2\gamma_k^2}\right) + 2\frac{\exp\left(-\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2\lambda_{k,\ell}^2R_K^2}\right)^{\left\lfloor\frac{\tilde{B}}{2\gamma_k}\right\rfloor}}{\exp\left(\frac{2\varepsilon^2}{\binom{n-k}{\ell}^2\lambda_{k,\ell}^2R_K^2}\right)-1}.
\end{aligned}
$$

□

# F DESCRIPTION OF MODELS, DATASETS AND EXPLANATION TASKS

We briefly sketched the datasets and models on which our cooperative games, used for the experiments, are built. Hence, we provide further details and sources to allow for reproducibility. Note that the LM, CNN, and SOUM are akin to (Fumagalli et al., 2023).

## F.1 Language Model (LM)

We used a pretrained sentiment analysis model for movie reviews. To be more specific, it s a variant of `DistilBert`, fine-tuned on the IMDB dataset, and its python version can be found in the *transformers* API (Wolf et al., 2020) at `https://huggingface.co/lvwerra/distilbert-imdb`. The explanation task is to explain the model's sentiment rating between $-1$ and 1 for randomly selected instances, where positive model outputs indicate positive sentiment. The features, which are words in this case, are removed on the token level, meaning that tokens of missing values are removed from the input sequence of words, shortening the sentence. Thus, a coalition within a given sentence is given by the sequence containing only the words associated with each each player of that coalition. The value function is given by the model's sentiment rating.

## F.2 Vision Transformer (ViT)

The ViT is, similar to the LM, a transformer model. Unlike the LM, the ViT operates on image patches instead of words. The python version of the underlying ViT model can be found in the *transformers* API at `https://huggingface.co/google/vit-base-patch32-384`. It originally consists of 144 32x32 pixel image patches, 12 patches for each column and row. In order to calculate the ground truth values exhaustively via brute force, we cluster smaller input patches together into 3x3 images containing 9 patches in total or into 4x4 images containing 16 patches in total. Patches of a cluster are jointly turned on and off depending on whether the cluster is part of the coalition or not. Players, represented by image patches, that are not present in a coalition are removed on the token level and their token is set to the empty token. The worth of a coalition is the model's predicted class probability for the class which has the highest probability for the grand coalition (the original image with no patches removed) and is therefore within $[0, 1]$.

## F.3 Convolutional Neural Network (CNN)

The next local explanation scenario is based on a ResNet18[2] model (He et al., 2016b) trained on ImageNet (Deng et al., 2009). The task is to explain the predicted class probability for randomly selected images from ImageNet (Deng et al., 2009). In order to obtain a player set, we use SLIC (Achanta et al., 2012) to merge single pixels to 14 super-pixels. Each super-pixel corresponds to a player in the resulting cooperative game, and a coalition of players entails the associated super-pixels. Absent super-pixel players are removed by setting the contained pixels to grey (mean-imputation). The worth of a coalition is given by the model's predicted class probability, using only the present super-pixels, for the predicted class of the full image with all super-pixels at hand.

## F.4 Sum Of Unanimity Models (SOUM)

We further consider synthetic cooperative games, for which the computation of the ground truth values is feasible within polynomial time. For a given player set $\mathcal{N}$ with $n$ many players, we draw $D = 50$ interaction subsets $S_1, \ldots, S_D \subseteq \mathcal{N}$ uniformly at random from the power set of $\mathcal{N}$. Next, we draw for each interaction subset $S_d$ a coefficient $c_d \in [0, 1]$ uniformly at random. The value function is simply constructed by defining

$$\nu(S) = \sum_{d=1}^{D} c_d \cdot [\![S_d \subseteq S]\!]$$

for all coalitions $S \subseteq \mathcal{N}$. We generate 50 instances of such synthetic games and average the approximation results. To our advantage, this construction yields a polynomial closed-form solution of the underlying CII values (Fumagalli et al., 2023), which allows us to use higher player numbers than in real-world explanation scenarios. For details of the CII computation we refer the interested reader to (Fumagalli et al., 2023).

[2] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html

# G FURTHER EMPIRICAL RESULTS

We conducted more experiments than shown in the main part but had to omit them due to space constraints. Besides the approximation curves, comparing SVARM-IQ's approximation quality for the SII, STI, and FSI against current baselines measured by the MSE and Prec@10, we present another type of visualization to demonstrate how SVARM-IQ's performance advantage aids in enriching explanations by including interaction effects.

## G.1 Further Results on the Approximation Quality

This section contains more detailed versions of the figures depicted in the main section. We compare the approximation quality of SVARM-IQ against baselines for the SII on the LM and ViT in Figure 6, for SII, STI, and FSI for CNN in Figure 7, and for SOUM in Figure 8.

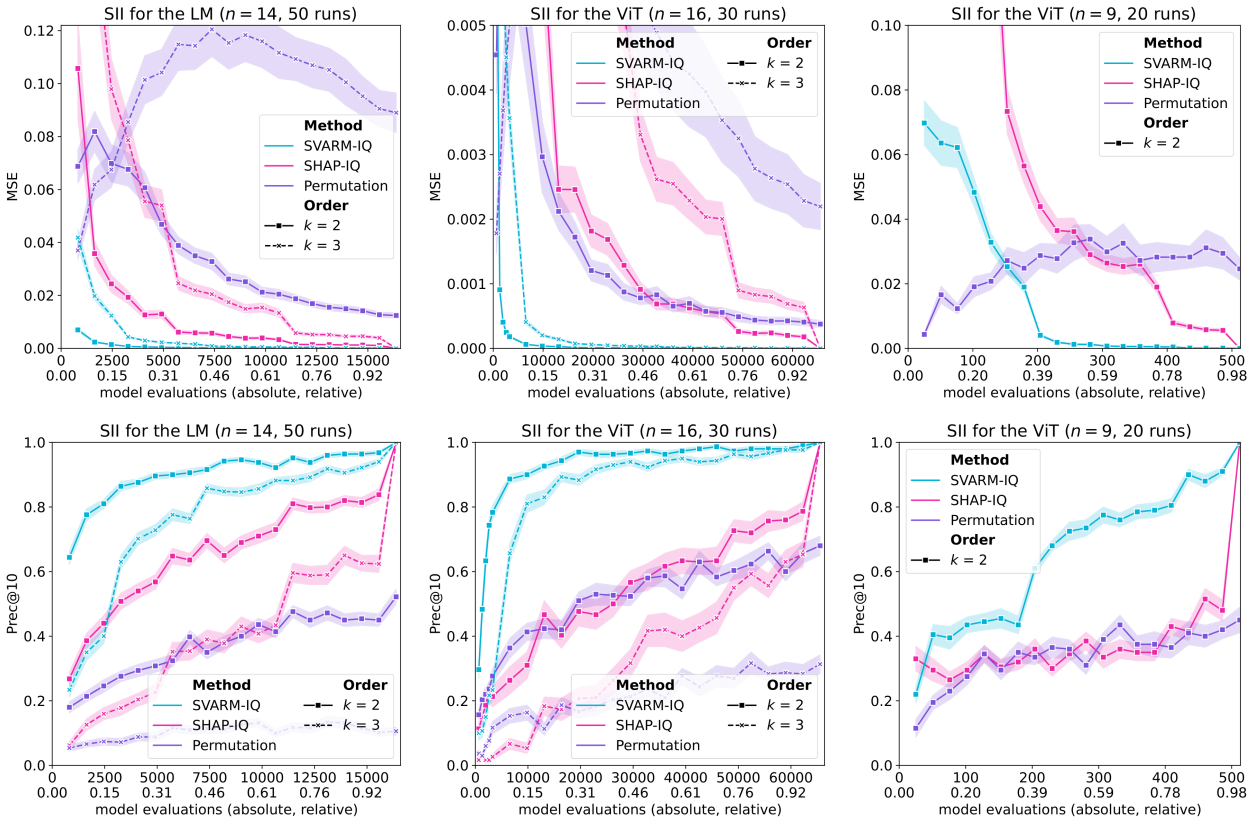



Figure 6: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines averaged over multiple runs for estimating the SII of order $k = 2, 3$ on the LM (first column, $n = 14$, 50 runs) and the ViT (second column, $n = 16$, 30 runs; second column, $n = 9$, 20 runs). The performance is measured by the MSE (first row) and Prec@10 (second row). The shaded bands represent the standard error over the number of performed runs.

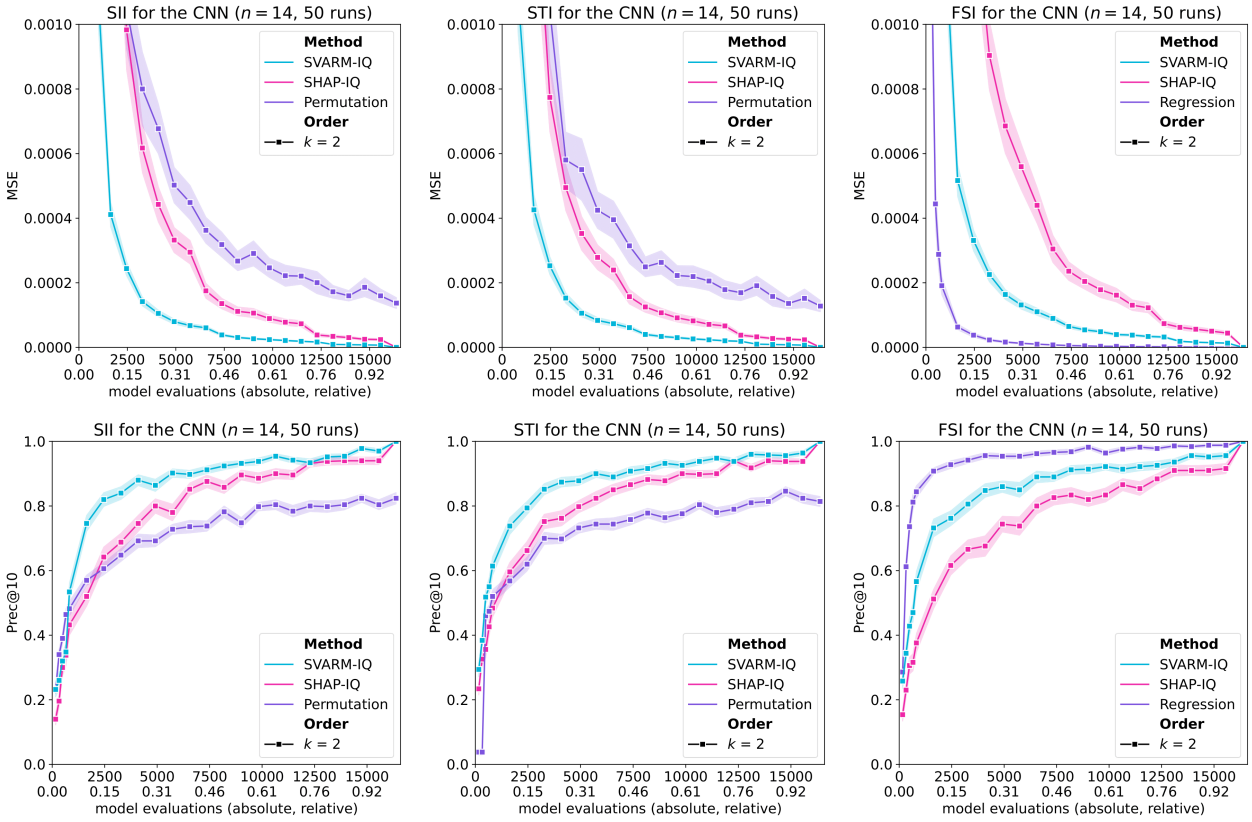



Figure 7: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines averaged over 50 runs on the CNN for estimating the SII (first column), STI (second column), and FSI (third column) of order $k = 2$ for $n = 14$. The performance is measured by the MSE (first row) and Prec@10 (second row). The shaded bands represent the standard error over the number of performed runs.

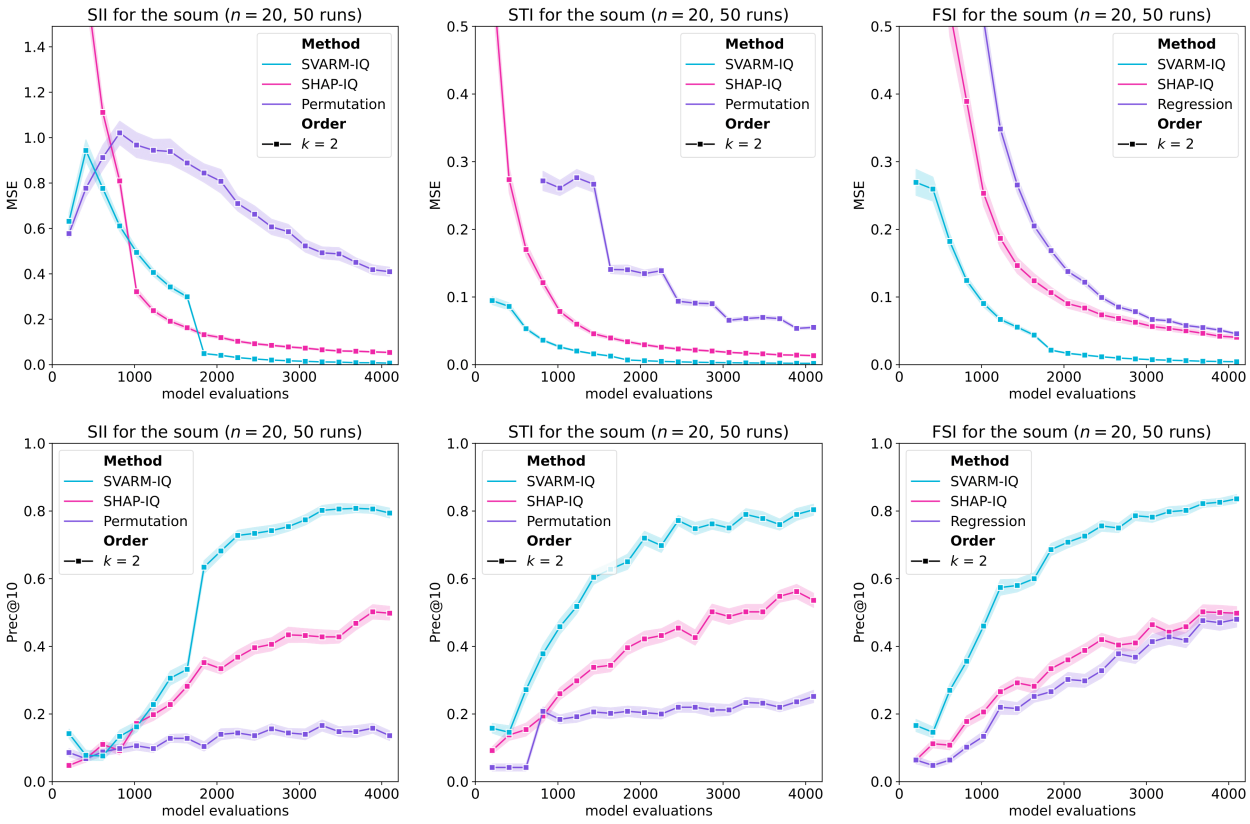



Figure 8: Approximation quality of SVARM-IQ (blue) compared to SHAP-IQ (pink) and permutation sampling (purple) baselines averaged over 50 runs on the SOUM for estimating the SII (first column), STI (second column), and FSI (third column) of order $k = 2$ for $n = 20$. The performance is measured by the MSE (first row) and Prec@10 (second row). The shaded bands represent the standard error over the number of performed runs.

### G.2 Further Examples of the Vision Transformer Case Study

In the following, we demonstrate how the inclusion of interaction besides attributions scores may enrich interpretability and how significantly SVARM-IQ contributes to more reliable explanations due to faster converging interaction estimates. First, we present in Figure 9 SVARM-IQ's estimates for our ViT scenario, which quantify the importance and interaction of image patches, revealing the insufficiency of sole importance scores and emphasizing the contribution of interaction scores for explaining class predictions for images. Second, we compare in Figure 10 attribution scores and interaction values estimated by SVARM-IQ and permutation sampling with the ground truth. Our results showcase that even with a relatively low number of model evaluations SVARM-IQ mirrors the ground truth almost perfectly, while the inaccurate estimates of its competitor pose the visible risk of misleading explanations, thus harming interpretability.

The obtained estimates for the labrador picture in Figure 9 (upper left) allow for a plausible explanation of the model's reasoning. The most important image patches, those which capture parts of the dogs' heads, share some interesting interaction. The three patches which contain at least one full eye, might be of high importance, but also exhibit strongly negative pairwise interaction. This gives us the insight that the addition of such a patch to an existing one contributes on average little to the predicted class probability in comparison to the increase that such a patch causes on its own, plausibly due to redundant information. In other words, it suffices for the vision transformer to see one patch containing eyes and further patches do not make it much more certain about its predicted class. On the other side, some patches containing different facial parts show highly positive interaction. For example, the teeth and the pair of eyes complement each other since each of them contains valuable information that is missing in the other patch. Considering only the importance scores and their ranking would have not led to this interpretation. Quite the opposite, practitioners would assume most patches to be of equal importance and overlook their insightful interplay.

The comparison of estimates with the ground truth in Figure 10 allows for a twofold conclusion. The estimates obtained by SVARM-IQ show barely any visible difference to the human eye. In fact, SVARM-IQ's approximation replicates the ground truth with only a fraction of the number of model evaluations that are necessary for its exact computation. Hence, it significantly lowers the computational burden for precise explanations. On the contrary, permutation sampling yields estimated importance and interaction scores which are afflicted with evident imprecision. Both, the strength and sign of interaction values are estimated with quite severe deviation for the two considered orders. Hence, the attempt to order the true interactions' strengths or identifying the most influential pairs becomes futile. This lack in approximation quality has the potential to misguide those who seek for explanations on why the model has predicted a certain class.

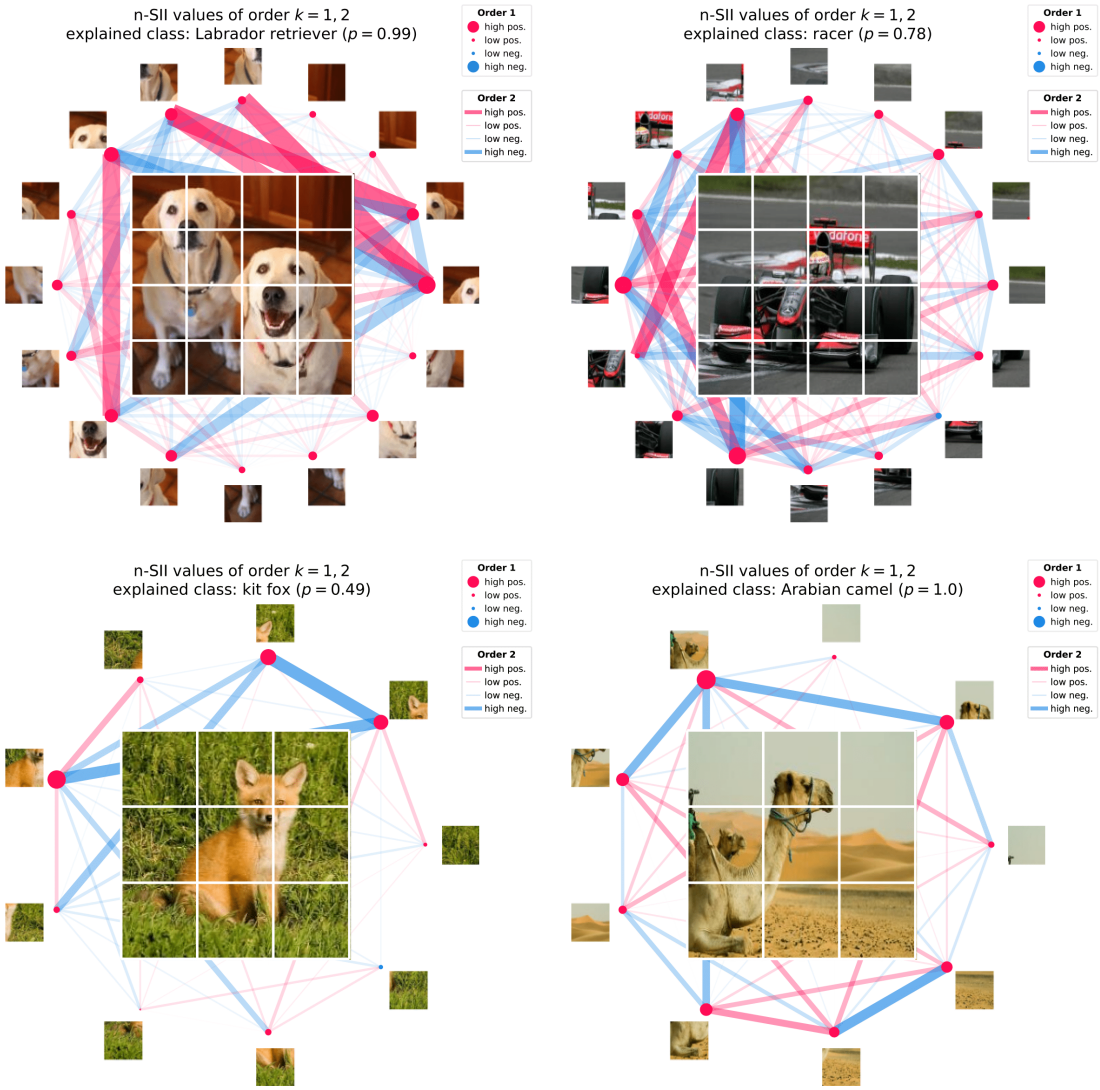



Figure 9: Computed n-SII values of order $k = 1, 2$ by SVARM-IQ for the predicted class probability of a ViT for selected images taken from ImageNet (Deng et al., 2009). The images are sliced into grids of multiple patches, $n = 16$ in the first row and $n = 9$ in the second row. The estimates are obtained after single computation runs given a budget of 10000 evaluations for $n = 16$ patches and 512 (GTV) for $n = 9$ patches.

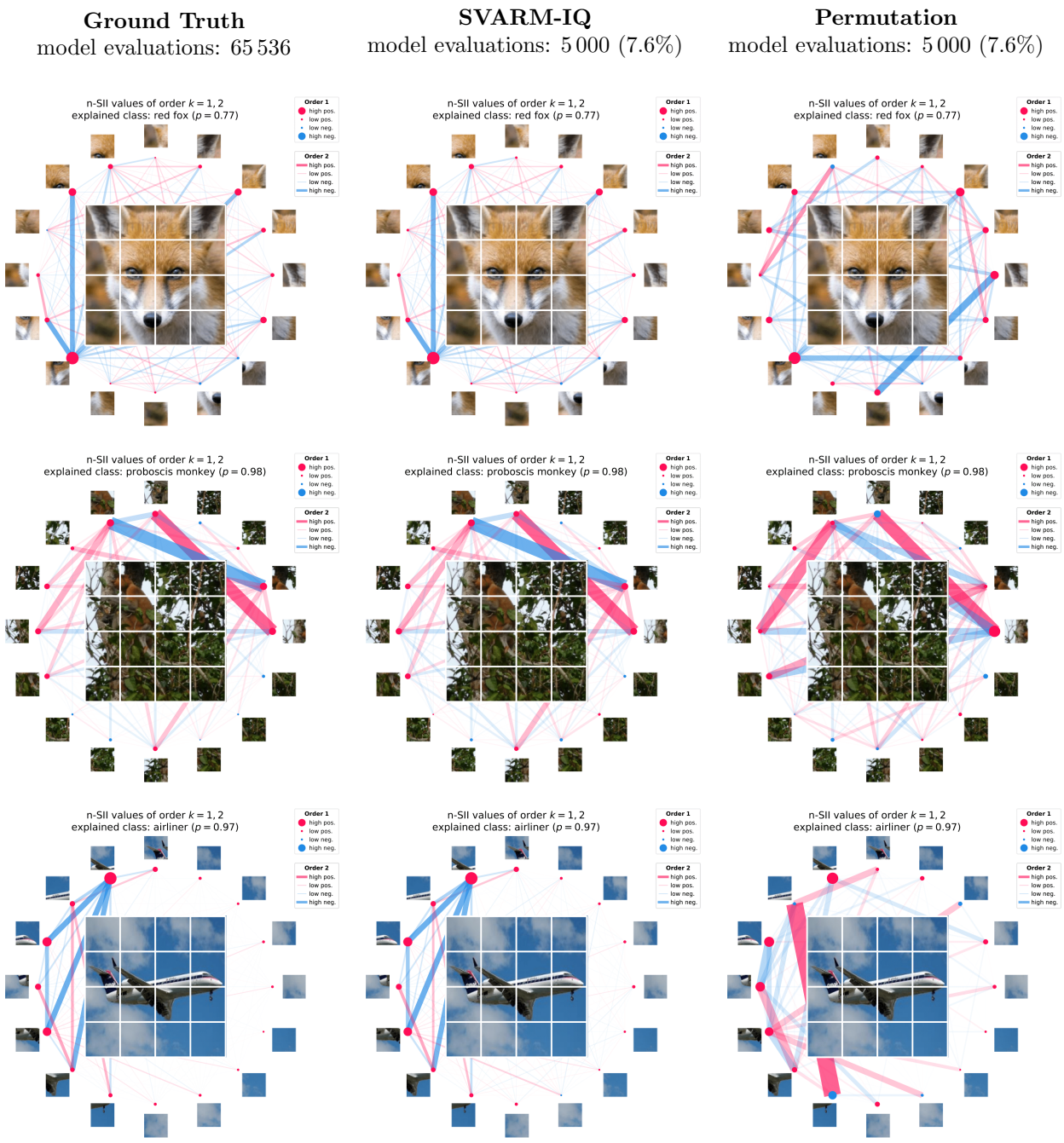



Figure 10: Row-wise comparison of ground-truth n-SII values of order $k = 1, 2$ for the predicted class probability of a ViT (first row) against n-SII values estimated by SVARM-IQ (second column) and permutation sampling (third row) with 5000 model evaluations. The pictures are taken from ImageNet (Deng et al., 2009) and sliced into a grid of 16 patches ($n = 16$).

# H HARDWARE DETAILS

This section contains the hardware details required to run and evaluate the empirical results. All experiments where developed and run on a single DELL XPS 15 9510 notebook with Windows 10 Education installed as the operating system. This laptop contains one 11th Gen Intel(R) Core(TM) i7-11800H clocking at 2.30GHz base frequency, 16.0 GB (15.7 GB usable) of RAM, and a NVIDIA GeForce RTX 3050 Ti Laptop GPU.

The model-function calls were pre-computed in around 10 hours on the graphics card. The evaluation of the approximation quality required around 50 hours of work on the CPU. In total, running the experiments took around 50 hours on a single core (no parallelization) and 10 hours on the graphics card.