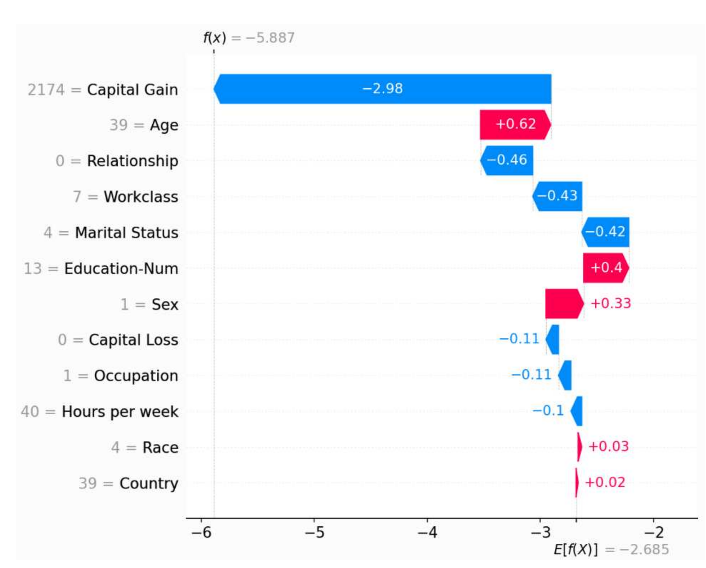
SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification

Patrick Kolpaczki[1], Maximilian Muschalik[2], Fabian Fumagalli[3], Barbara Hammer[3], and Eyke Hüllermeier[2]

1) Paderborn University, Germany 2) LMU Munich, Germany 3) Bielefeld University, Germany

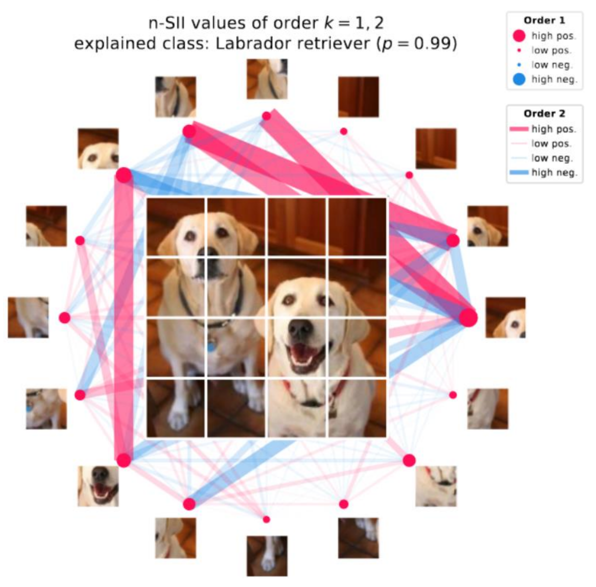
Motivation: Feature Explanations

Interactions enrich Shapley-based feature explanations:



Shapley Feature Attribution

Additively decomposes prediction f(x) on a data point among the features



Pairwise Feature Interaction

Quantifies the synergy effect on f(x) of two features being present

Shapley-based Interaction Indices

- **Player set** $\mathcal{N} = \{1,\dots,n\}$ → Features, datapoints, neurons, base learners etc.
- **Value function** $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ with $\nu(\emptyset) = 0$ → Predicted value, generalization performance

Definition: Shapley Value (Shapley, 1953)

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

Marginal contribution $\Delta_i(S)$: **Increase** in collective benefit when i joins S.

- Unique solution to fulfill desirable axioms: Symmetry, Additivity, Null-Property, Efficiency
- Computational effort scales **exponentially** with n : 2^n coalitions in total

Definition: Cardinal Interaction Indices (Fujimoto et al., 2006)

For pairs i, j :

$$I_{i,j} = \sum_{S \subseteq \mathcal{N}\setminus\{i,j\}} \lambda_{2,|S|} \cdot \underbrace{[\nu(S \cup \{i,j\}) - \nu(S \cup \{i\}) - \nu(S \cup \{j\}) + \nu(S)]}_{=\Delta_{i,j}(S)}$$

For subsets K of order k:

$$I_K = \sum_{S \subseteq \mathcal{N}\setminus K} \lambda_{k,|S|} \cdot \Delta_K(S) \quad \text{with} \quad \Delta_K(S) := \sum_{W \subseteq K} (-1)^{|K|-|W|} \cdot \nu(S \cup W)$$

Discrete derivative $\Delta_K(S)$: **Synergy effect** of K at the presence of S.

The weights $\lambda_{k,|S|}$ define the specific Interaction Index:

- **Shapley Interaction Index (SII):** $\lambda^{\text{SII}}_{k,|S|} = \frac{1}{(n-k+1) \cdot \binom{n-k}{|S|}}$
- **Shapley-Taylor Interaction Index (STI):** $\lambda^{\text{STI}}_{k,|S|} = \frac{k}{n \cdot \binom{n-1}{|S|}}$

And many more:

- Faithful-Shapley Interaction Index (FSI)
- Banzhaf Interaction Index (BII)

Fixed-budget approximation problem:

- Given cooperative game (N, ν) with unknown Interaction scores I_K for all $K \subseteq N$ of order k
- Budget B : Allowed number of evaluations of ν (bottleneck due to model access)
  Model evaluations (inference, retraining) pose bottleneck on runtime rather than arithmetic operations
- Minimize mean squared error (MSE) averaged over all estimates $\hat{I}_K$: $\frac{1}{\binom{n}{k}} \sum_{K \subseteq \mathcal{N}} \mathbb{E}\left[\left(\hat{I}_K - I_K\right)^2\right]$

Contribution

Novel approximation algorithm for all Shapley-based Interactions:

- Powerful combination of stratified representation + update mechanism
- Applicable to **all orders** (pairs, triples, etc.) and **indices** (SII, STI, etc.) simultaneously
- Novel theoretical guarantees & state-of-the-art empirical performance

✓ **Model-agnostic** / domain-independent
→ *Applicable to any model and data, and even outside of explainability and ML*

✓ No hyperpameters → *No fine-tuning*

✓ Estimates available at any time → *Budget can be cut and extended arbitrarily*

Approximation Algorithm: SVARM-IQ

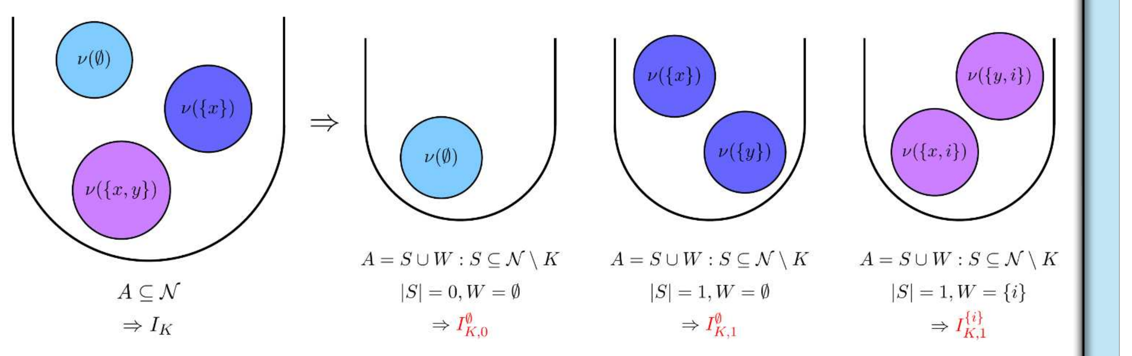
Stratified Representation

$$I_K = \sum_{\ell=0}^{n-k} \binom{n-\ell}{k} \underbrace{\lambda_{k,\ell} \sum_{W \subseteq K} (-1)^{k-|W|} \cdot I^W_{K,\ell}}_{=\, I_{K,\ell}\text{: Weighted average of strata with } |S|=\ell} \quad \text{with} \quad I^W_{K,\ell} := \underbrace{\frac{1}{\binom{n-k}{\ell}} \sum_{\substack{S \subseteq \mathcal{N}\setminus K \\ |S|=\ell}} \nu(S \cup W)}_{\text{Stratum: Average worth of coalitions of size } \ell+|W|\text{, containing only } W \text{ out of } K}$$

Double Stratification of I_K into $I_{K,\ell}$ and $I^W_{K,\ell}$ *for fixed k*

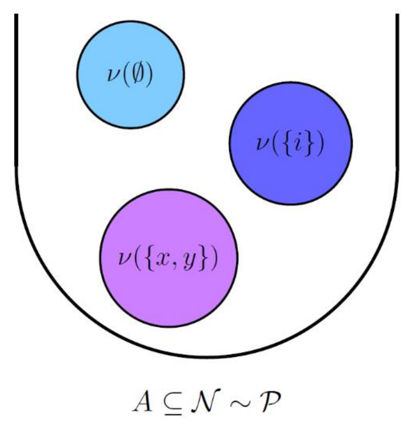
- Partition powerset into $2^k\binom{n}{k}(n-k+1)$ strata
- Coalitions are grouped by size $\ell + |W|$ and intersection $W = A \cap K$
- Strata are more homogeneous than base population
- Enables enhanced update mechanism
- Maintain estimates $\hat{I}^W_{K,\ell}$

Example for $K = \{i,j\}$



Sampling Coalitions and Udapting all Estimates

- Calculate all border strata exactly with coalitions of size $0, \dots, b, n-b, \dots, n$
- Perform warmup on inner strata $b+1, \dots n-b-1$ to initialize all estimates
- Repeat with remaining budget $\tilde{B}$:
  - Draw coalition size $s \in \{b+1, \dots, n-b-1\} \sim \tilde{P}(s)$
  - Draw coalition A of size s uniformly at random and evaluate $\nu(A)$
  - For all K: Update $\hat{I}^W_{K,\ell}$ with $W = A \cap K$ and $\ell = |A| - |W|$



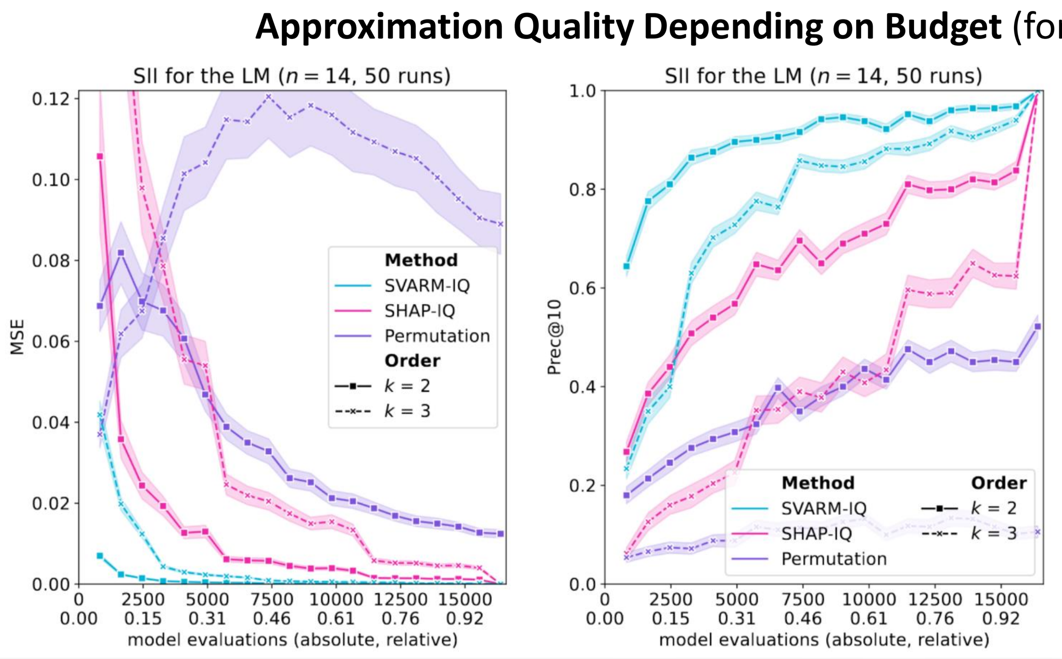
Theorem 4.2 & Corollary 4.3: Variance and MSE

For any K of order k, the variance and MSE of the estimate $\hat{I}_K$ returned by SVARM-IQ is bounded by

$$\mathbb{E}\left[\left(\hat{I}_K - I_K\right)^2\right] = \mathbb{V}\left[\hat{I}_K\right] \leq \frac{\gamma_k}{\tilde{B}} \sum_{W \subseteq K} \sum_{\ell=0}^{n-k} \binom{n-k}{\ell}^2 \lambda^2_{k,\ell} \sigma^2_{K,\ell,W} .$$

- *With stratum variances $\sigma^{+\,2}_{K,\ell,W} = \mathbb{V}[\nu(A \cup W)]$ for $A \subseteq \mathcal{N} \setminus K$ with $|A| = \ell$ drawn u.a.r.*
- $\gamma_k := 2(n-1)^2$ for $k = 2$ and $\gamma_k = n^{k-1}(n-k+1)^2$ for $k \geq 3$

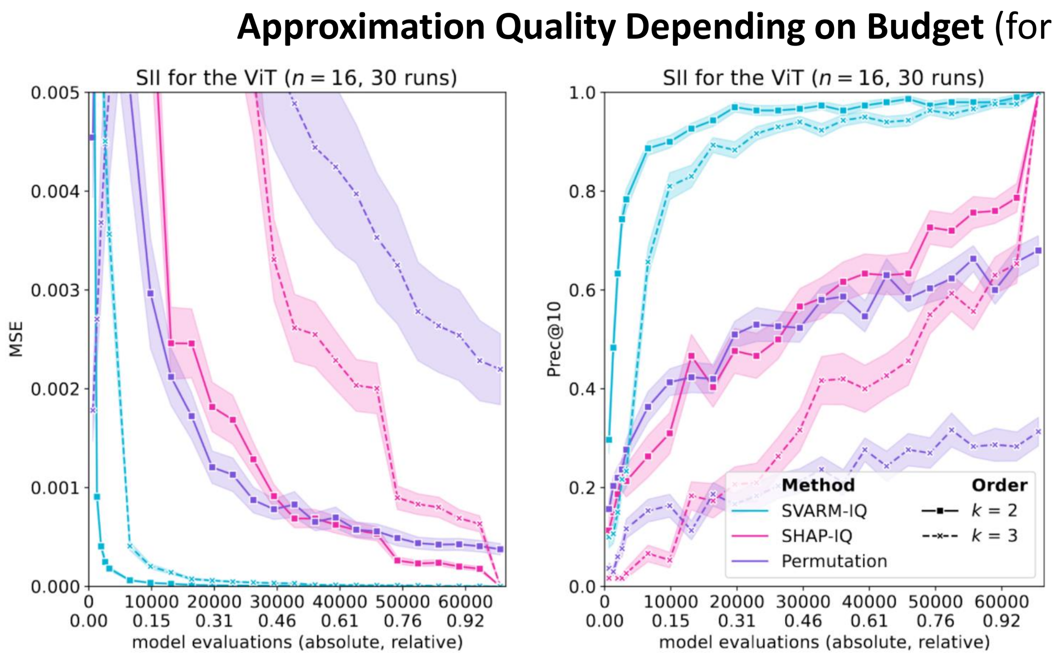
Approximation Quality Depending on Budget (for Language Model)



Setup for local explanations:

- Pretrained sentiment analysis model DistilBert
- Sentences of 14 words length
- Features (words) are removed on token level
- Value function: Model's sentiment rating in [-1,1]

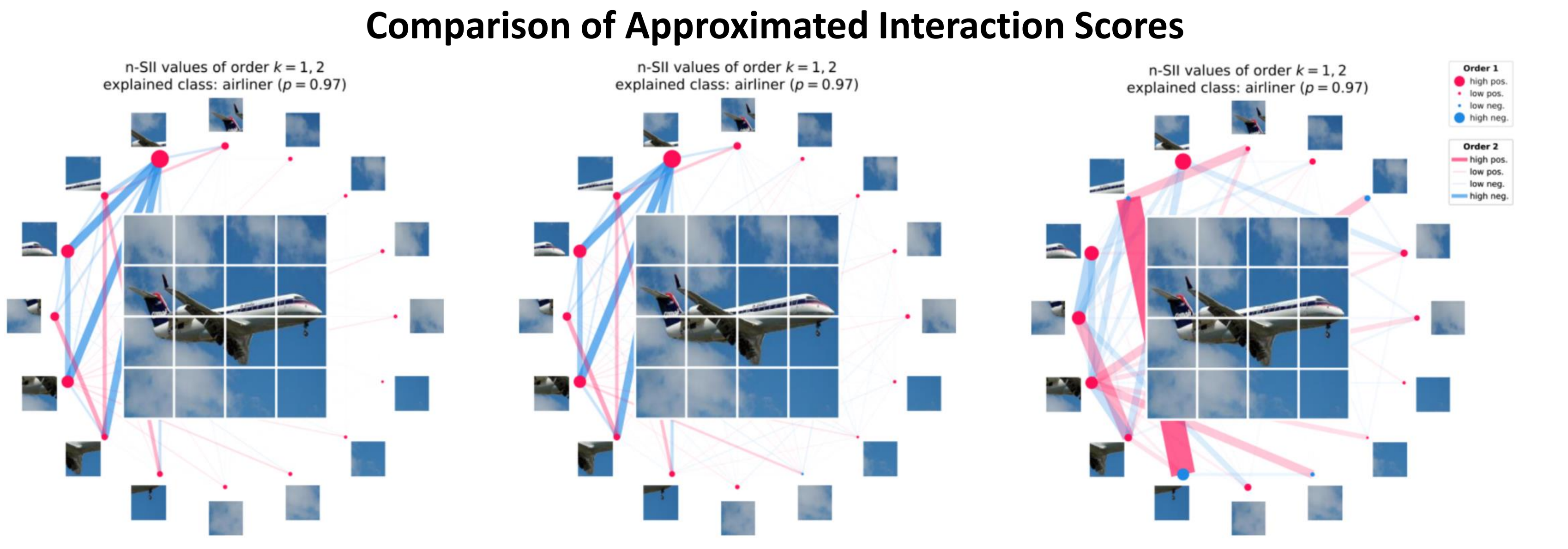
Approximation Quality Depending on Budget (for Vision Transformer)



Setup for local explanations:

- Vision transformer model operating on image patches
- ImageNet pictures sliced into 4x4 grid of 16 patches
- Features (patches) are removed by turning them off
- Value function: Predicted class probability for class predicted with full image

Comparison of Approximated Interaction Scores



Ground Truth | SVARM-IQ | Permutation Sampling (Tsai et al.2023)

- Players: 16 grid patches of an ImageNet picture
- Value function: Predicted class probability of a Vision Transformer for class predicted with full image
- Approximators are run with a budget of 5000 samples (7.6% of 65,536 required model evaluations for exact calculation)

References

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. (2023). SHAP-IQ: Unified Approximation of any-Order Shapley Interactions. In *Proceedings of Advances in Neuron Processing Systems.*

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.*

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.*

Kolpaczki, P., Bengs, V., Muschalik, M., Hüllermeier, E. (2024). Approximating the Shapley Value without Marginal Contributions. In *Proceedings of AAAI Conference on Artifical Intelligence.*

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.*

Sundararajan, M., Dhamdhere, K., and Agrarwal, A. (2020). The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning.*

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

Ministry of Culture and Science of the State of North Rhine-Westphalia

dataninja.nrw

arXiv Paper

Code Package