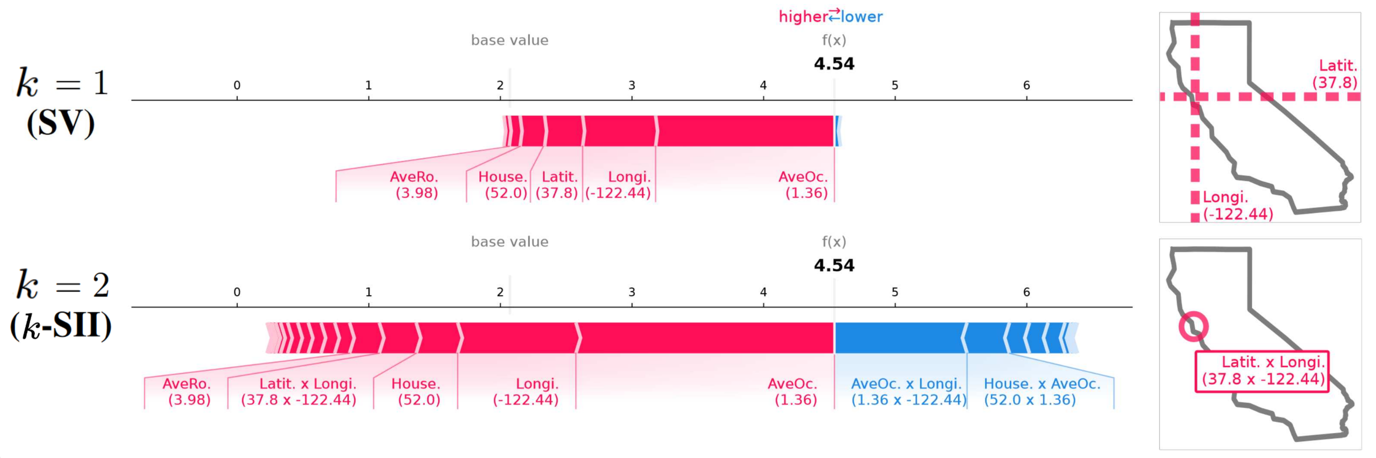
KernelSHAP-IQ: Weighted Least Square Optimization for Shapley Interactions

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

ICML International Conference On Machine Learning

Interaction Example: Explaining Property Prices

Potential Question: "Does the *location* of my property affect its price?"



Background

Shapley Value (SV) [1]:

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

value function (e.g. model prediction)

marginal contribution (increase in collective benefit when i joins T)

Shapley Interaction Index (SII) [2]:

$$\phi^{SII}(S) := \sum_{T \subseteq N \setminus S} \frac{(n-s-t)! \cdot t!}{(n-s+1)!} \underbrace{\sum_{L \subseteq S} (-1)^{s-l} \nu(T \cup L)}_{=:\Delta_S(T)}$$

discrete derivative (synergy effect of S in the presence of T)

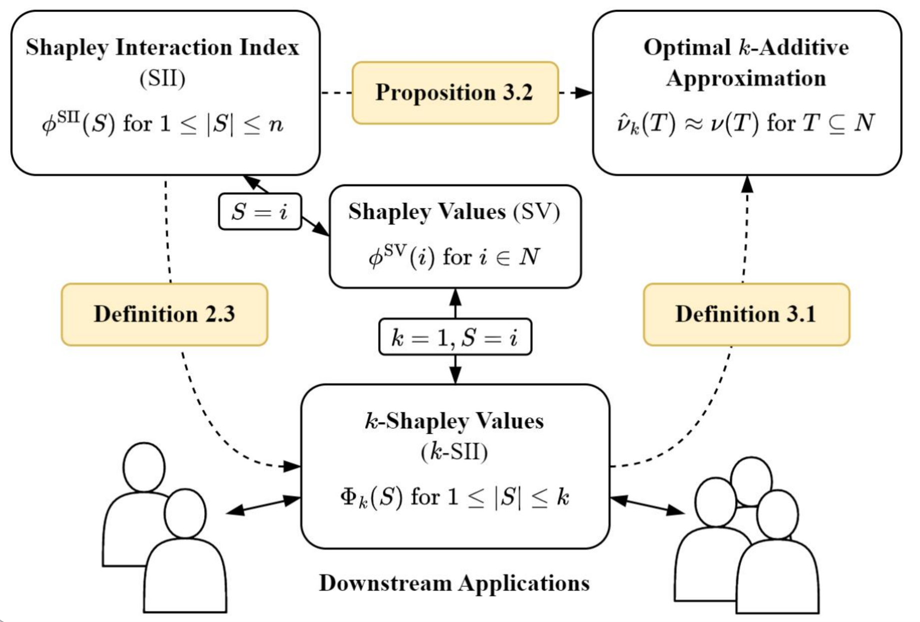
k-Shapley Values (k-SII) [3]:

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

Bernoulli numbers

Notation: Player set N; $i,j \in N$; $S \subseteq N$ **Convention:** $s := |S|$ or $n := |N|$

Link between SV, SII and k-SII



Definition 3.1 (k-additive Approximation)

$$\hat{\nu}_k(T) := \sum_{S \subseteq T}^{1 \le |S| \le k} \Phi_k(S)$$

- SII is an order-independent basis index
- k-SII is well-suited for interpretability by adjusting explanation order k

References

[1] Shapley, L. S. (1953). A Value for n-Person Games. In Contributions to the Theory of Games, Volume II, pages 307–318. Princeton University Press.
[2] Grabisch, M. and Roubens, M. (1999). An Axiomatic Approach to the Concept of Interaction Among Players in Cooperative Games. Int. J. Game Theory, 28(4):547–565
[3] Bordt, S. and von Luxburg, U. (2023). From Shapley Values to Generalized Additive Models and Back. In AISTATS'23, pp. 709–745.
[4] Lundberg et al., (2017). A Unified Approach to Interpreting Model Predictions. In NeurIPS'17.
[5] Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. (2023). SHAP-IQ: Unified Approximation of any-order Shapley Interactions. In NeurIPS'23.
[6] Guilherme, D. P., Duarte L. T., and Grabisch, M. (2023). A k-additive Choquet Integral-Based Approach to Approximate the SHAP Values for Local Interpretability in Machine Learning. In Artif. Ingell., 325:104014
[7] Kolpaczki, P., Bengs, V., Muschalik, M., and Hüllermeier, E. (2024). Approximating the Shapley Value without Marginal Contributions. In AAAI'24, pp. 13246–13255
[8] Muschalik, M., Fumagalli, F., Hammer, B., and Hüllermeier, E. (2024). Beyond TreeSHAP: Efficient Computation of Any-Order Shapley Interactions for Tree Ensembles. In AAAI'24, pp. 14388–14396
[9] Kolpaczki, P., Muschalik, M., Fumagalli, F, Hammer, B, and Hüllermeier, E. (2024). SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification. In AISTATS'23, pp. 3520–3528

Contribution

TLDR: We present a novel **least-square representation** for the **Shapley Interaction Index** (SII) [2] and present a kernel-based estimator called **KernelSHAP-IQ** akin to KernelSHAP [4] for the Shapley value (SV) [1].

KernelSHAP-IQ

KernelSHAP [4] utilizes a weighted least-square representation:

$$\phi^{SV} = \arg\min_{\phi \in \mathbb{R}^n} \sum_{T \subseteq N}^{0<|T|<n} \mu_1(t) \left[\nu(T) - \sum_{i \in T} \phi(i)\right]^2 \text{ s.t. } \sum_{i \in N} \phi(i) = \nu(N)$$

A new **weighted least-square representation** for the SII:

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

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

SV residuals; fit for $\hat{\nu}_2$

with:

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

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

SII-specific weights

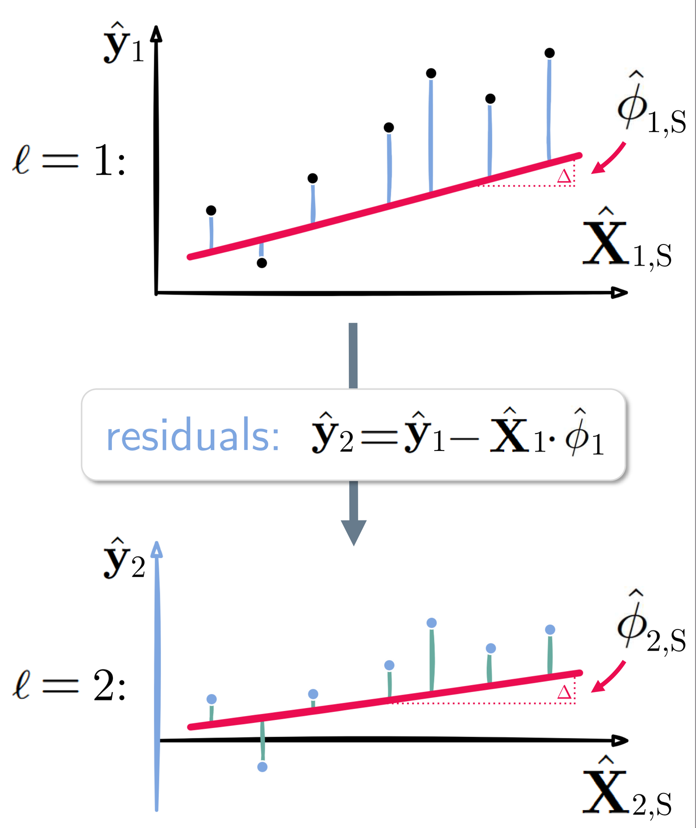
KernelSHAP-IQ is a recursive optimization (simplified for 2-SII):

for order $l = 1, 2$:

- sample subsets from budget $\{T_i\}_i, \{w_T\}_T \leftarrow \textsc{Sample}(b)$
- evaluate game/model $\hat{\mathbf{y}}_1 \leftarrow [\nu(T_1), \ldots, \nu(T_b)]^T$
  - adjust weights per order $\hat{\mathbf{X}}_\ell, \hat{\mathbf{W}}_\ell^* \leftarrow \textsc{Weight}(\ell, \ldots)$
  - solve the regression $\hat{\phi}_\ell \leftarrow \textsc{SolveWLS}(\hat{\mathbf{X}}_\ell, \hat{\mathbf{y}}_\ell, \hat{\mathbf{W}}_\ell^*)$
  - compute residuals $\hat{\mathbf{y}}_{\ell+1} \leftarrow \hat{\mathbf{y}}_\ell - \hat{\mathbf{X}}_\ell \cdot \hat{\phi}_\ell$
- aggregate into k-SII (optional) $\hat{\Phi}_2 \leftarrow \textsc{AggregateSII}(\hat{\phi}_1, \hat{\phi}_2)$

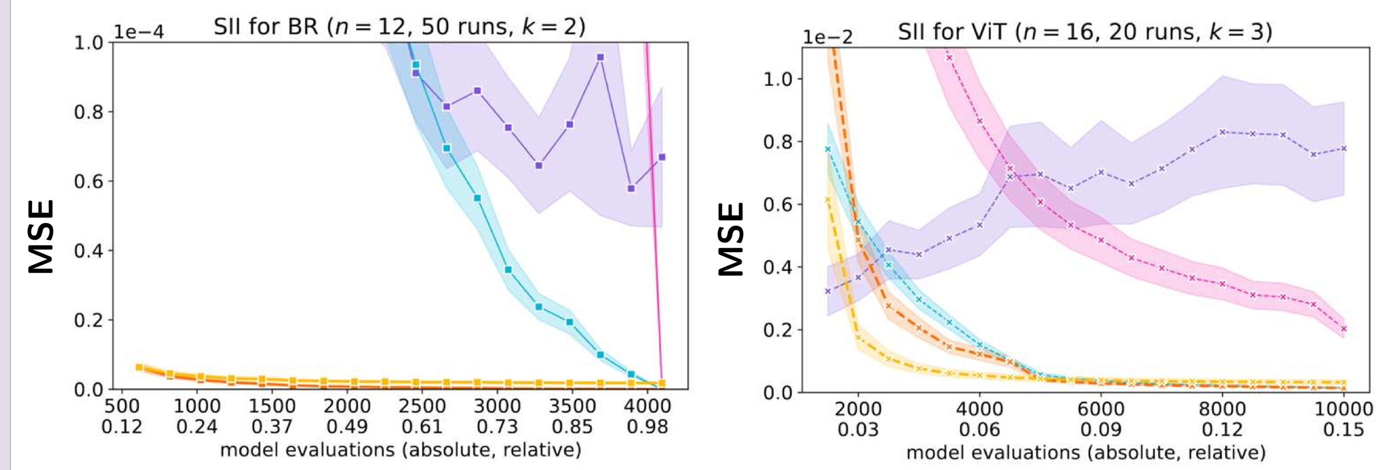
➢ order $l = 1$: KernelSHAP
➢ order $l \geq 2$: KernelSHAP-IQ

illustration for interaction S:

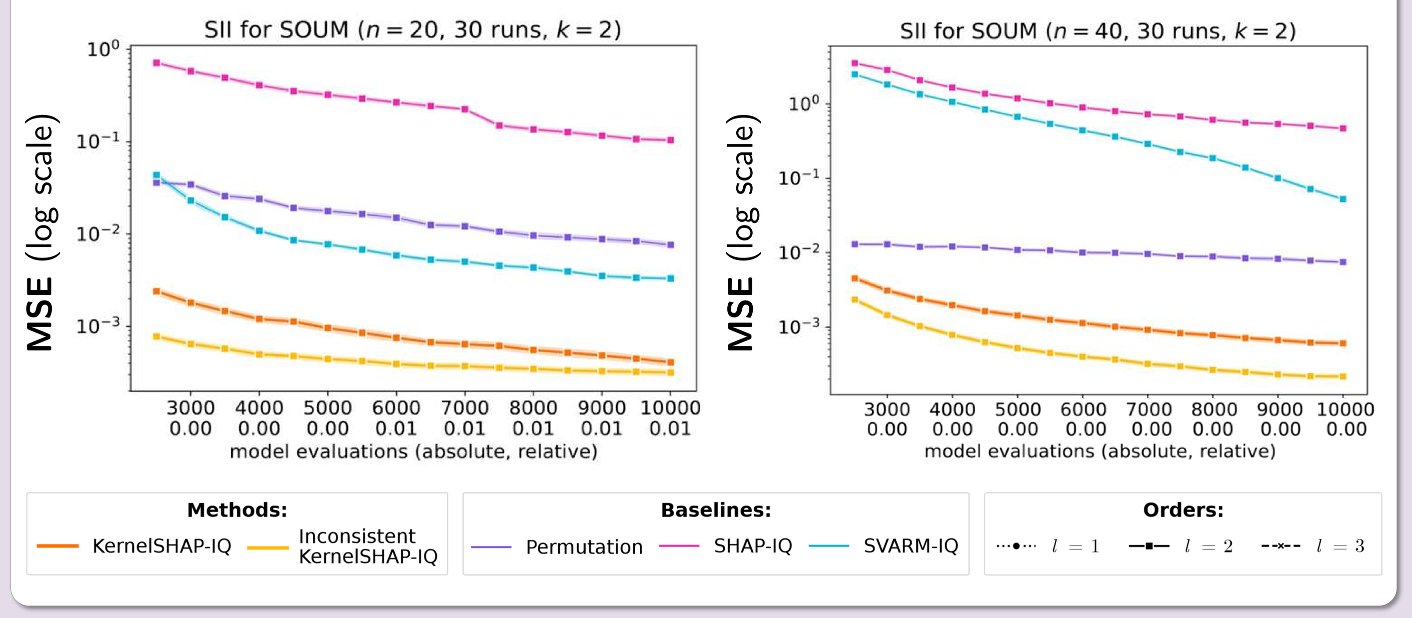


Empirical Results

➢ XAI benchmark for bike regression (BR) and vision transformer (ViT):

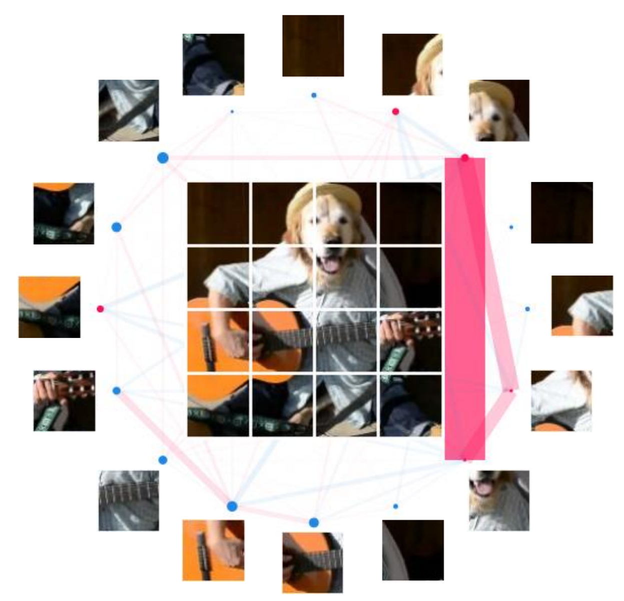


➢ sum of unanimity models (SOUMs) with high number of features n:



Network Plot for a ViT:

- model predicts the class ***golden retriever*** with probability $p_{\max} = 0.203$
- the highest attribution score is the **second order interaction** between the ***head*** and the ***snout***



Open-Source Implementation

- **KernelSHAP-IQ** is available for python

```
pip install shapiq
```

- shapiq includes **18** game theoretic concepts including SV, SII, k-SII, BV, ...
- around **20** approximators and explainers including SHAP-IQ [5], SVARM-IQ [7,9], KernelSHAP [4], k-add SHAP [6], TreeSHAP-IQ [8], ...
- **plot and interpret interactions** with different visualization techniques

1) Bielefeld University, Bielefeld, Germany 2) LMU Munich, Munich, Germany 3) MCML, Munich, Germany 4) Paderborn University, Paderborn, Germany

Funded by Deutsche Forschungsgemeinschaft German Research Foundation
Ministry of Culture and Science of the State of North Rhine-Westphalia