Approximating the Shapley Value without Marginal Contributions

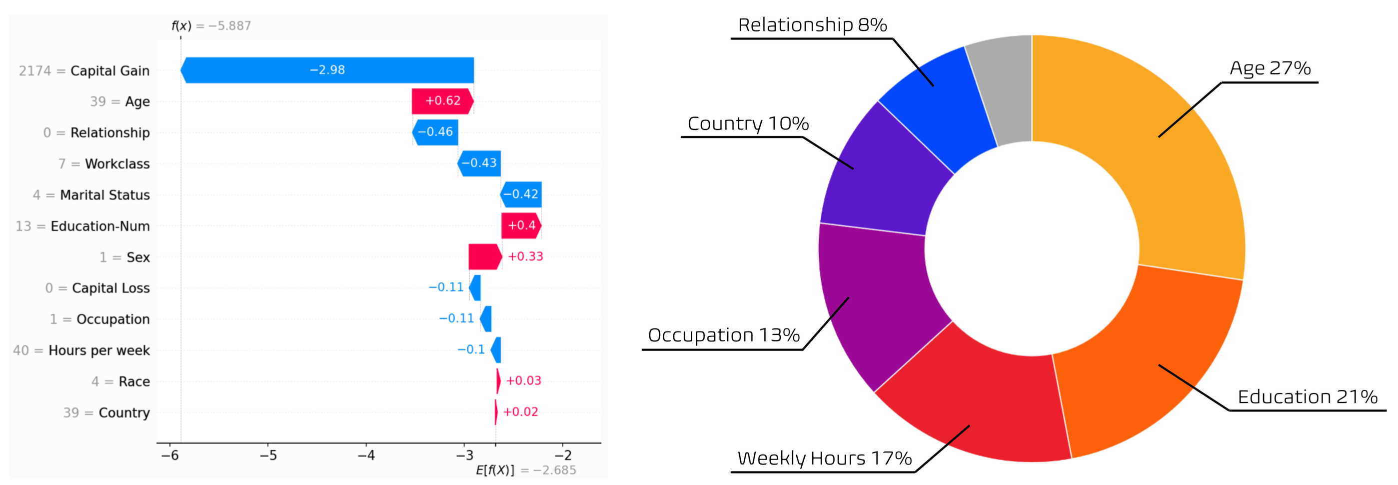
Patrick Kolpaczki[1], Viktor Bengs[2],
Maximilian Muschalik[2], and Eyke Hüllermeier[2]

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

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20–27, 2024 | VANCOUVER, CANADA

Motivation: Additive Feature Explanations

Feature explanations quantify each feature's contribution to a numerical effect:



Local: Feature Attribution

Quantifies contribution to the prediction value of a specified data point.

Global: Feature Importance

Quantifies contribution to the model's generalization performance.

Shapley Value

- **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: Efficiency, Symmetry, Additivity, Null-Property
- Computational effort scales **exponentially** with n : 2^n coalitions in total

Fixed-budget approximation problem:

- Given cooperative game (N, ν) with unknown Shapley values $\phi_1, \dots, \phi_n$
- Budget T : 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 players for estimates $\hat{\phi}_1, \dots, \hat{\phi}_n$:

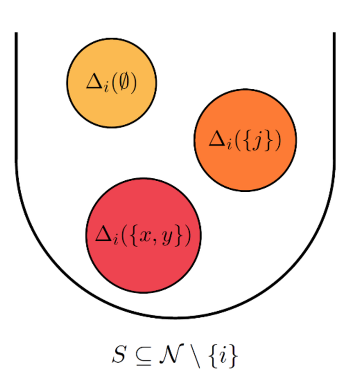
$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right]$$

Approximation by sampling marginal contributions:

- Weights form a well-defined probability distribution: $\sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} = 1$

→ The Shapley value is the expected marginal contribution: $\phi_i = \mathbb{E}[\Delta_i(S)]$

- Obtain $\hat{\phi}_i$ by sampling marginal contributions according to weights
- One separate approximation problem for each player



Problem: *Notion of marginal contributions is inefficient!*

One update of $\hat{\phi}_i$ with $\Delta_i(S) = \nu(S \cup \{i\}) - \nu(S)$ costs **2 budget tokens**

Contribution

We propose a novel approximation algorithm for the Shapley value with:

- New combination of stratified representation + update mechanism
- Novel strong theoretical guarantees
- State-of-the-art empirical performance

- **Desirable properties:**
  - ✓ Model-agnostic / domain-independent
    → Applicable for data valuation, neuron importance, etc. and even outside of ML
  - ✓ No hyperparameters → *No fine-tuning*
  - ✓ Estimates available at any time → *Budget can be cut and extended arbitrarily*
  - ✓ Uncertainty-aware → *Allows construction of confidence intervals*

Approximation Algorithm

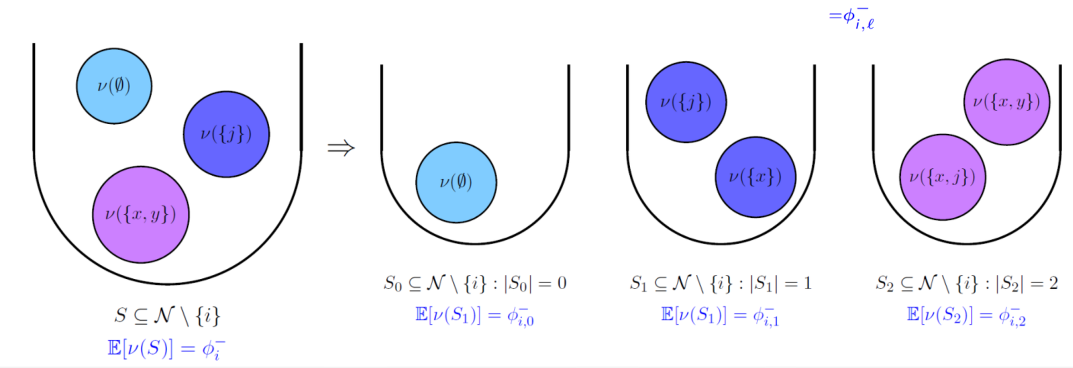
Stratified Representation

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

Stratification of ϕ_i^+ and ϕ_i^-:

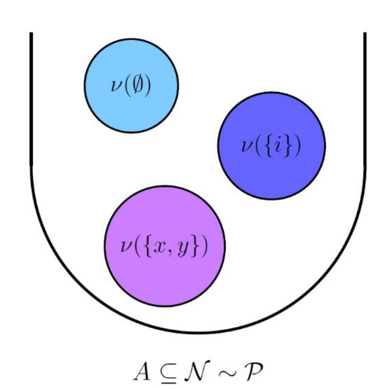
$$\phi_i^- = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot \nu(S) \quad \Rightarrow \quad \phi_i^- = \frac{1}{n} \sum_{\ell=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{\ell}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \{i\} \\ |S| = \ell}} \nu(S)}_{=\phi_{i,\ell}^-}$$

- Partitions the marbles into n many strata
- Strata are grouped by size
- More homogeneous than base population
- Enables enhanced update mechanism
- Maintain estimates $\hat{\phi}_{i,\ell}^+$, $\hat{\phi}_{i,\ell}^-$



Stratified Shapley Value Approximation without Requesting Marginals (Stratified SVARM)

- Calculate all strata exactly with coalitions of size $0, 1, \dots, n-1, n$
- Perform warmup to initialize all estimates: $W \in O(n \log n)$
- Repeat with remaining budget $\bar{T} = T - W$:
  - Draw coalition size $s \in \{2, \dots, n-2\} \sim \tilde{P}(s)$
  - Draw coalition of size s uniformly at random and evaluate $\nu(A)$
  - Update $\phi_{i,s-1}^+$ for all $i \in A$
  - Update $\phi_{i,s}^-$ for all $i \notin A$



Theorem 6. & Corollary 2. Variance and MSE

The variance and MSE of any estimate $\hat{\phi}_i$ returned by Stratified SVARM is bounded by

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

with stratum variances ${\sigma_{i,\ell}^+}^2 = \mathbb{V}[\nu(A \cup \{i\})]$ and ${\sigma_{i,\ell}^-}^2 = \mathbb{V}[\nu(A)]$ for $A \subseteq \mathcal{N} \setminus \{i\}$ with $|A| = \ell$ drawn u.a.r.

Theorem 5. Unbiasedness

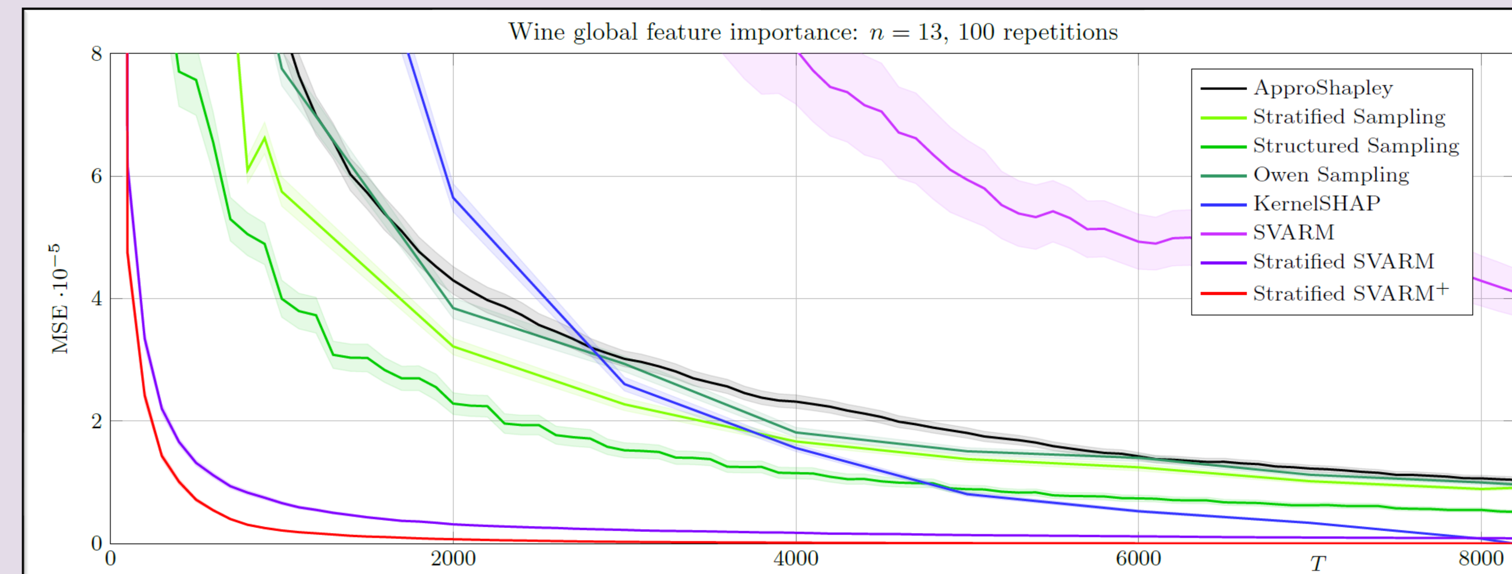
The estimate $\hat{\phi}_i$ of any $i \in \mathcal{N}$ returned by Stratified SVARM is unbiased, i.e., $\mathbb{E}\left[\hat{\phi}_i\right] = \phi_i \; \forall i \in \mathcal{N}$.

Theorem 7. PAC bound

For any estimate $\hat{\phi}_i$ returned by Stratified SVARM and $\varepsilon > 0$ holds

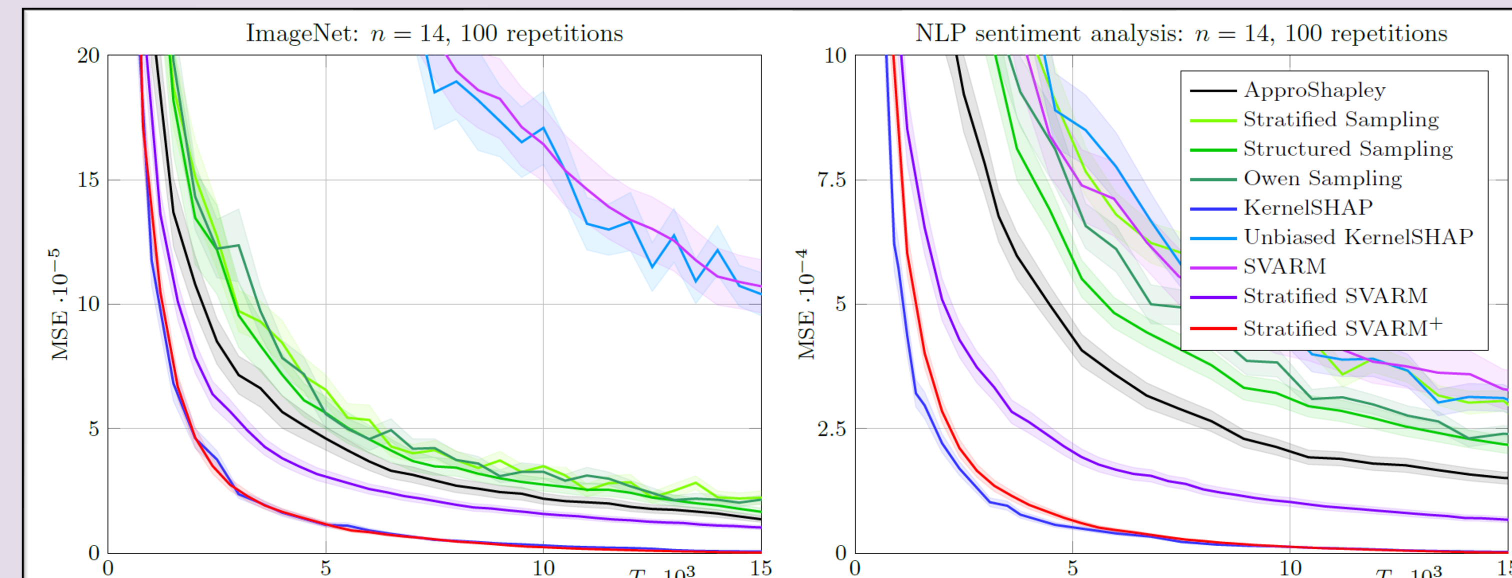
$$\mathbb{P}\left(|\hat{\phi}_i - \phi_i| \geq \varepsilon\right) \leq \frac{2 \log n}{\varepsilon^2 n \bar{T}} \sum_{\ell=2}^{n-2} {\sigma_{i,\ell-1}^+}^2 + {\sigma_{i,\ell}^-}^2 .$$

Empirical Evaluation



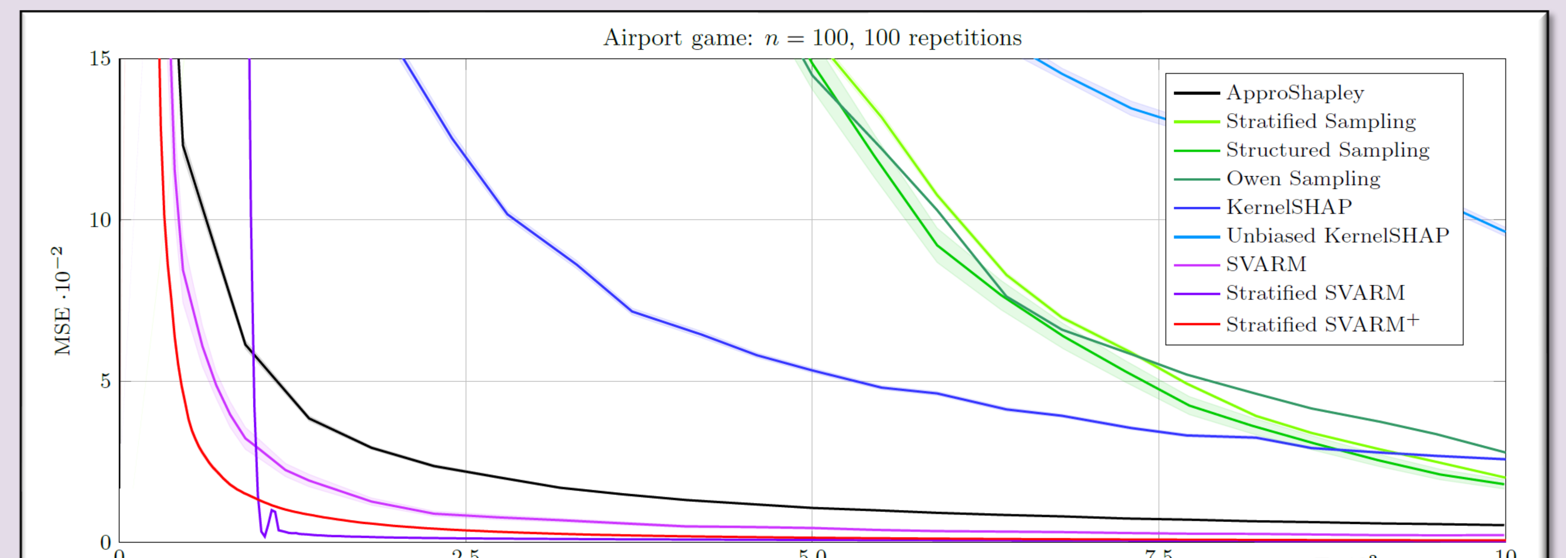
Global explanation task on real-world data: Divide prediction accuracy among features

- **Players:** 13 tabular features used for classifying origin of wine by random forest
- Feature removal: Retrain model on training data with only present coalitions
- **Value function:** Classification accuracy of random forest measured on test data



Local explanation tasks on real-world data: Divide predicted class probability among features

- **Players:** 14 semantic tiles (ImageNet), 14 tokenized words (NLP sentiment analysis)
- Pretrained model h predicts probability distribution over class labels
- Feature removal: Impute value by mean (numeric feature) or mode (categorical feature)
- Predicted class probability of label c using feature subset S : $h_S(c)$
- Predicted class label with full feature set: c^*
- **Value function:** $\nu(S) = h_S(c^*) - h_\emptyset(c^*)$



Synthetic airport game: Divide runway maintenance costs among airplanes

- **Players:** 100 airplanes of different size
- Each airplane has its own size c_i
- Size of biggest airplane determines costs
- **Value function:** $\nu(S) = \max_{i \in S} c_i$
- Infeasible naive exact calculation
- Structure elicits closed-form polynomial solution
  (but unknown to approximation algorithm)
- High degree of non-additivity

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

dataninja.nrw