Estimating Shapley Values for Explainable AI via Richer Model Approximations

R. Teal Witter, Yurong Liu,
Christopher Musco

Shapley Values

A (very) popular approach for explaining AI

shap Public

Watch 241 Fork 3.4k Star 24.6k

ACM Digital Library
https://dl.acm.org › doi

A unified approach to interpreting model predictions

by SM Lundberg · 2017 · Cited by 43101 — We present a unified framework for interpreting predictions, **SHAP** (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a ...

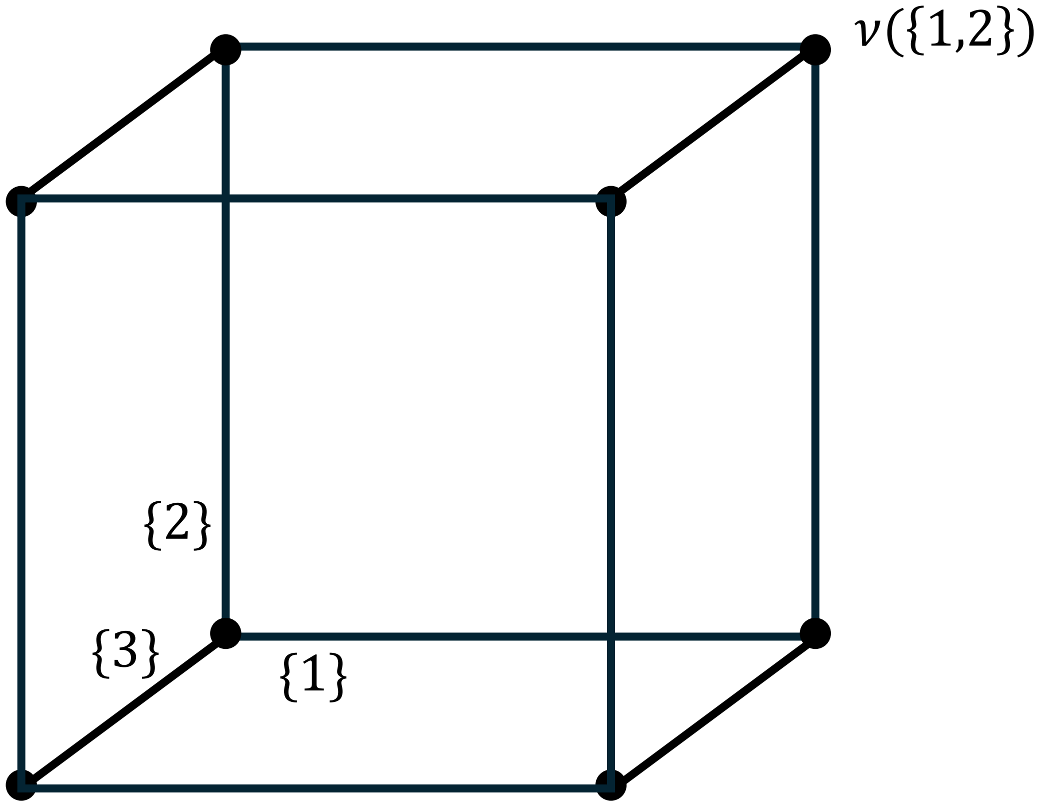
Value Functions

Notation: $[d] = \{1,2,\dots,d\}$

$$v: 2^{[d]} \to \mathbb{R}$$

Examples:

- $v(S)$ = Predictions when only given features in S
- $v(S)$ = Loss when only given features in S
- $v(S)$ = Predictions when only given points in S
- $v(S)$ = Loss when only given points in S

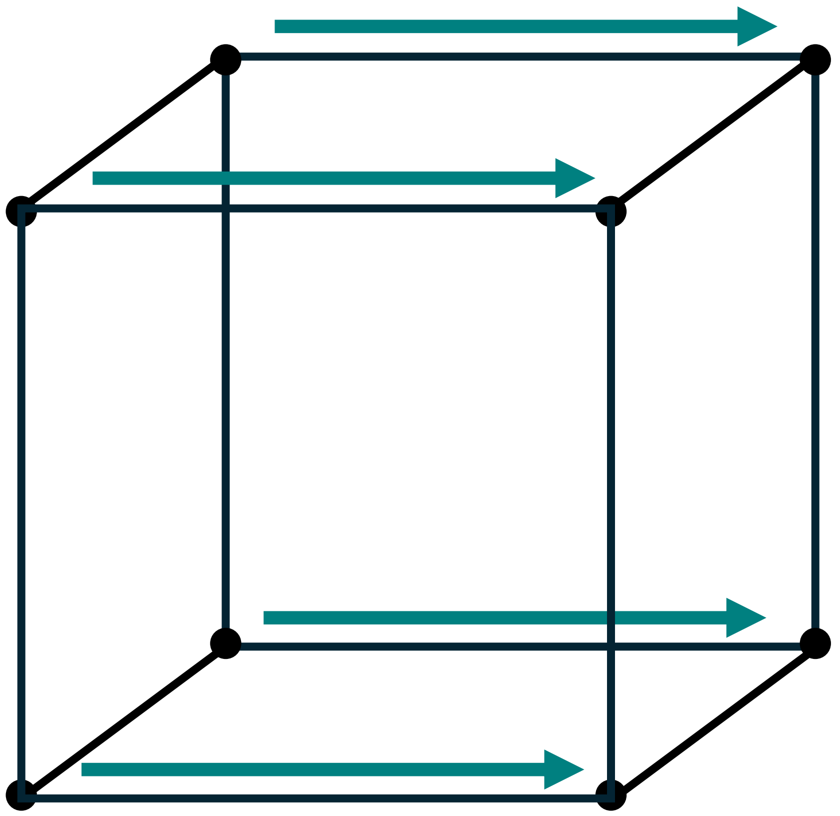


Shapley Values

$$\phi_i(v) = \mathbb{E}_S[v(S \cup \{i\}) - v(S)]$$

$$= \sum_{S \subseteq [d] \setminus \{i\}} [v(S \cup \{i\}) - v(S)]\, p_{|S|}$$

Challenge: 2^d subsets to evaluate!



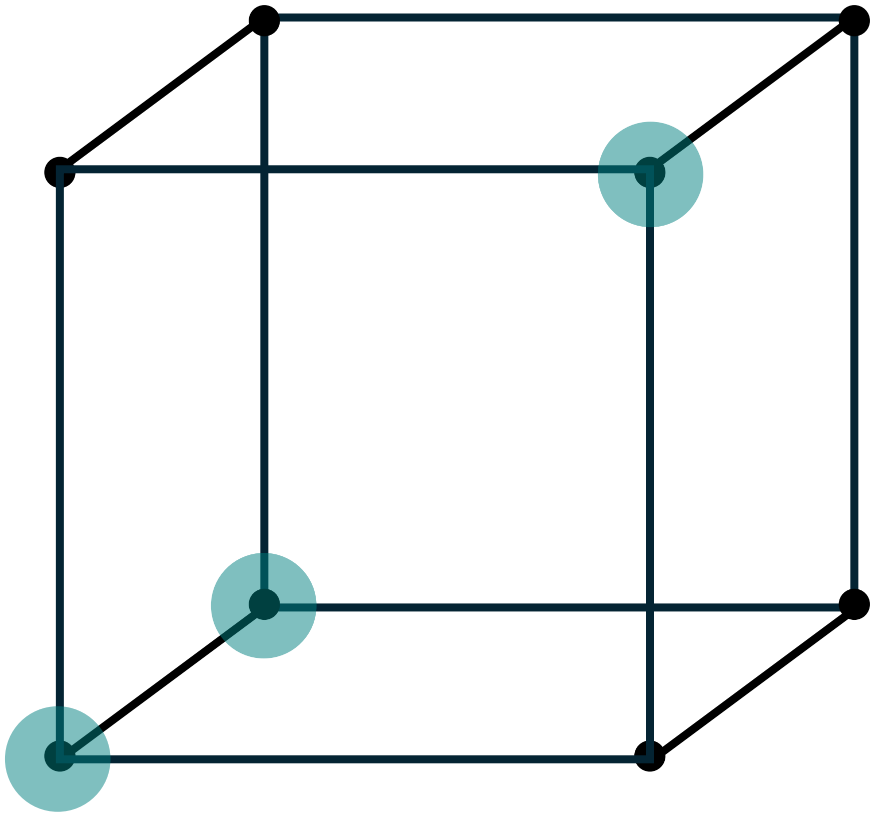
Kernel SHAP

Sample m subsets $S_1, \ldots, S_m \subseteq [d]$

Fit a linear function $f: 2^{[d]} \to \mathbb{R}$ on samples

Return $\phi_1(f), \ldots, \phi_d(f)$

Justification: When $m = 2^d, \phi_i(f) = \phi_i(v)$

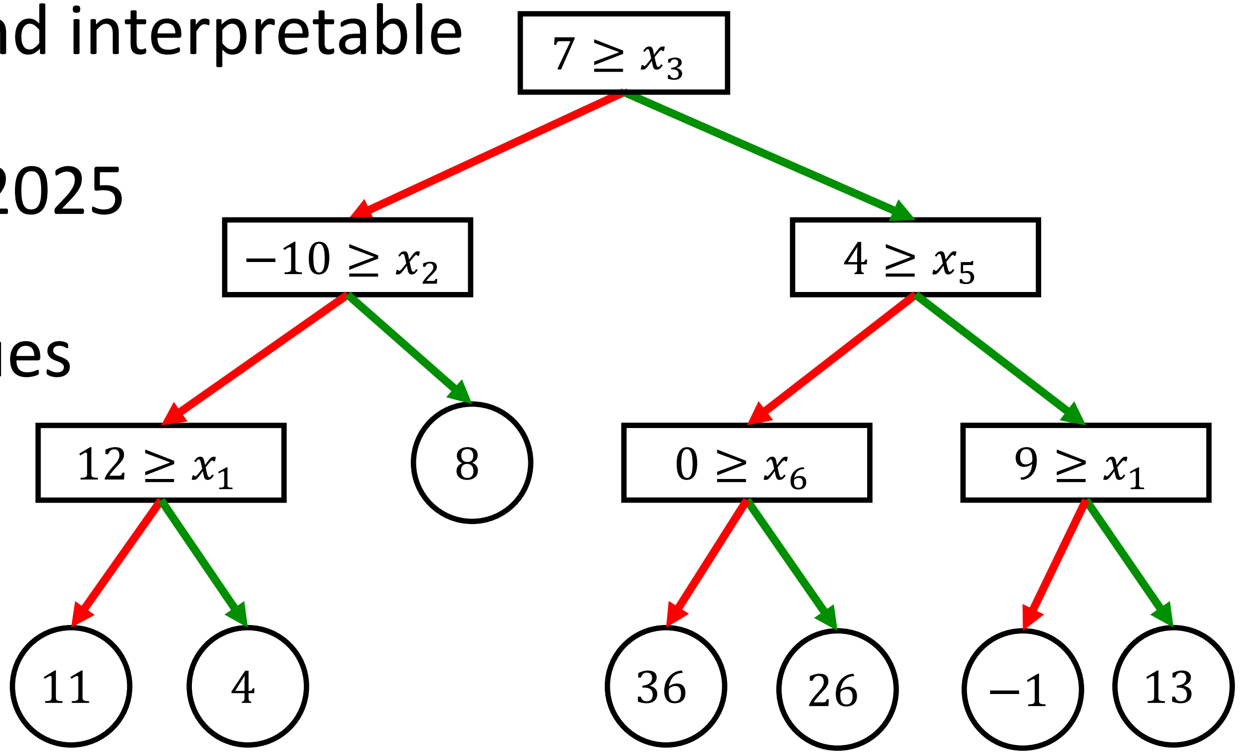


Guiding Question

What if we used an even more **expressive** function f to fit v?

Gradient Boosted Trees

- Efficient (*highly* optimized) and interpretable
- SOTA on tabular data 2015 - 2025
- Easy to compute Shapley values



Regression MSR

Sample subsets, **Fit** gradient boosted trees

Challenge: Even with all samples, $\phi_i(f) \neq \phi_i(\nu)$

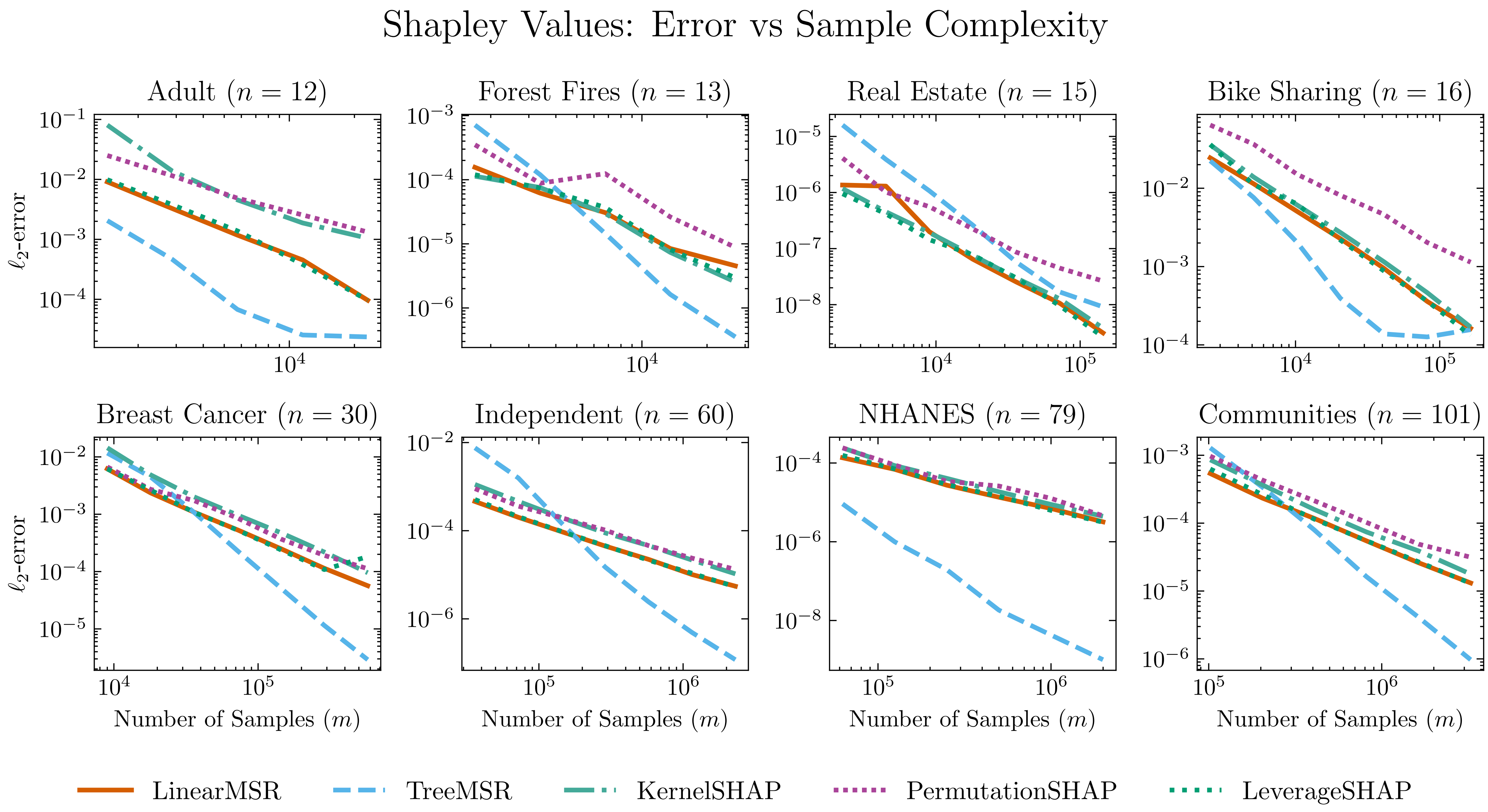
Return $\phi_i(f) + \hat{\phi}_i(\nu - f)$

Justification: With all samples, $\hat{\phi}_i(\nu - f) = \phi_i(\nu - f)$ and so

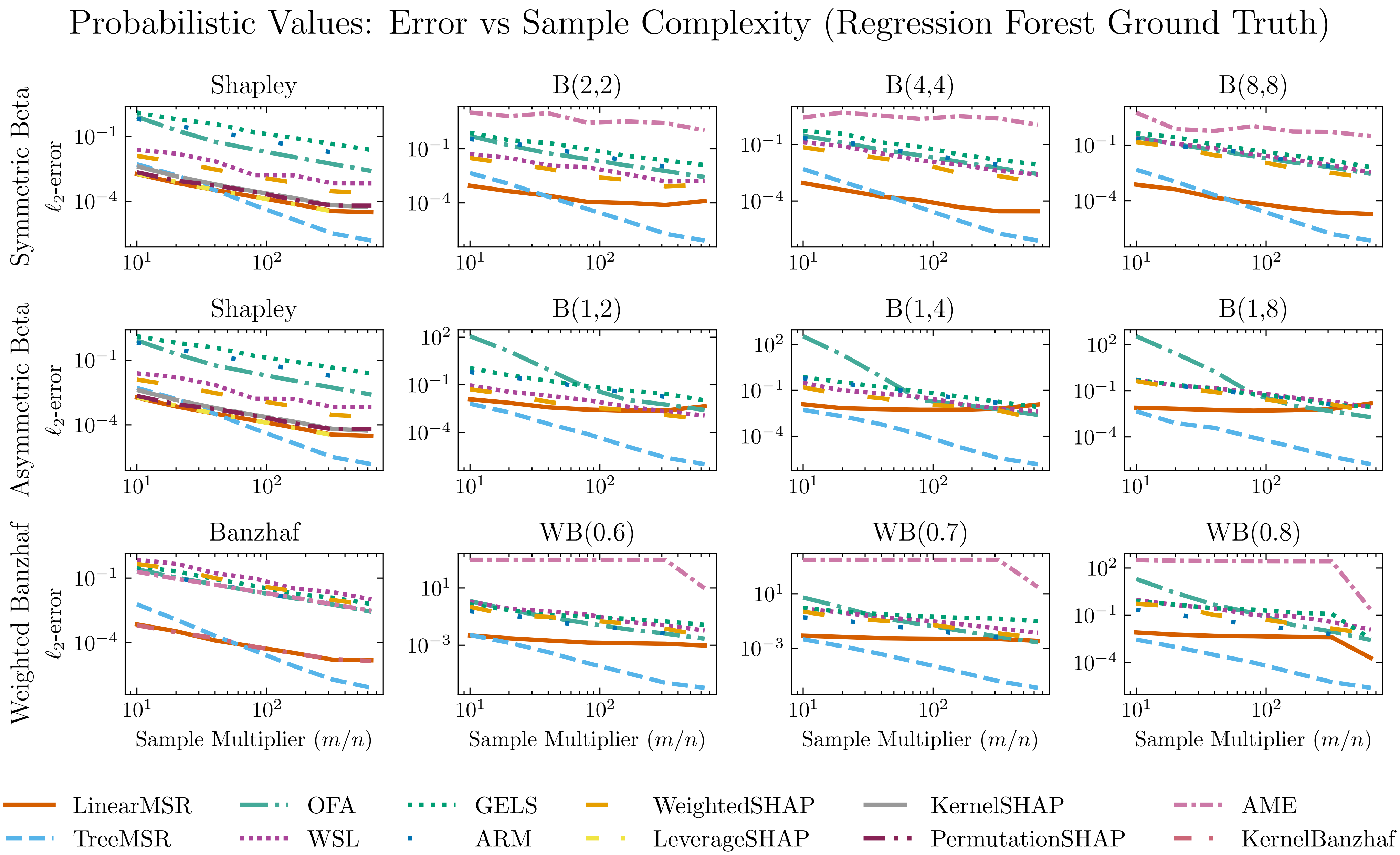
$$\phi_i(f) + \phi_i(\nu - f) = \phi_i(f + \nu - f) = \phi_i(\nu)$$

Shapley Value MSE

	Adult	Forest Fires	Real Estate	Bike Sharing	Breast Cancer	Independent	NHANES	Communities	Mean
LinearMSR									
Mean	1.18×10^{-3}	3.07×10^{-5}	2.00×10^{-7}	5.07×10^{-3}	1.09×10^{-3}	9.49×10^{-5}	2.73×10^{-5}	1.17×10^{-4}	9.51×10^{-4}
1st Quartile	2.99×10^{-4}	4.72×10^{-7}	2.46×10^{-8}	9.72×10^{-4}	2.14×10^{-4}	5.38×10^{-5}	2.18×10^{-10}	4.76×10^{-5}	1.98×10^{-4}
2nd Quartile	7.67×10^{-4}	2.75×10^{-6}	5.91×10^{-8}	2.85×10^{-3}	1.02×10^{-3}	6.64×10^{-5}	2.49×10^{-6}	9.46×10^{-5}	6.00×10^{-4}
3rd Quartile	1.52×10^{-3}	6.00×10^{-6}	1.82×10^{-7}	6.37×10^{-3}	1.71×10^{-3}	1.06×10^{-4}	2.88×10^{-5}	1.45×10^{-4}	1.24×10^{-3}
TreeMSR									
Mean	6.77×10^{-5}	1.45×10^{-5}	1.07×10^{-6}	2.04×10^{-3}	1.08×10^{-3}	1.47×10^{-4}	1.95×10^{-7}	7.93×10^{-5}	4.29×10^{-4}
1st Quartile	1.79×10^{-5}	1.32×10^{-6}	9.50×10^{-8}	6.23×10^{-4}	2.37×10^{-4}	2.40×10^{-5}	2.99×10^{-10}	1.78×10^{-5}	1.15×10^{-4}
2nd Quartile	4.12×10^{-5}	3.55×10^{-6}	1.97×10^{-7}	1.28×10^{-3}	5.51×10^{-4}	8.20×10^{-5}	8.89×10^{-10}	3.58×10^{-5}	2.50×10^{-4}
3rd Quartile	9.03×10^{-5}	1.01×10^{-5}	1.40×10^{-6}	2.44×10^{-3}	1.23×10^{-3}	1.70×10^{-4}	3.91×10^{-9}	5.70×10^{-5}	5.00×10^{-4}
KernelSHAP									
Mean	4.55×10^{-3}	2.98×10^{-5}	1.93×10^{-7}	6.12×10^{-3}	2.01×10^{-3}	1.97×10^{-4}	4.08×10^{-5}	1.59×10^{-4}	1.64×10^{-3}
1st Quartile	5.14×10^{-4}	3.08×10^{-7}	1.04×10^{-9}	1.40×10^{-3}	6.87×10^{-4}	1.09×10^{-4}	1.60×10^{-16}	7.03×10^{-5}	3.47×10^{-4}
2nd Quartile	8.59×10^{-4}	3.05×10^{-6}	3.50×10^{-8}	4.00×10^{-3}	1.89×10^{-3}	1.64×10^{-4}	3.10×10^{-6}	1.27×10^{-4}	8.81×10^{-4}
3rd Quartile	2.84×10^{-3}	7.30×10^{-6}	1.59×10^{-7}	7.91×10^{-3}	2.98×10^{-3}	2.80×10^{-4}	3.97×10^{-5}	2.25×10^{-4}	1.79×10^{-3}
PermutationSHAP									
Mean	4.86×10^{-3}	1.25×10^{-4}	5.58×10^{-7}	1.51×10^{-2}	1.73×10^{-3}	1.96×10^{-4}	3.43×10^{-5}	2.14×10^{-4}	2.78×10^{-3}
1st Quartile	1.65×10^{-3}	8.54×10^{-7}	3.64×10^{-9}	3.13×10^{-3}	2.97×10^{-4}	6.96×10^{-5}	1.60×10^{-16}	5.87×10^{-5}	6.50×10^{-4}
2nd Quartile	3.84×10^{-3}	4.83×10^{-6}	4.90×10^{-8}	5.97×10^{-3}	1.05×10^{-3}	1.70×10^{-4}	2.10×10^{-6}	1.61×10^{-4}	1.40×10^{-3}
3rd Quartile	7.68×10^{-3}	1.52×10^{-5}	2.69×10^{-7}	1.92×10^{-2}	1.97×10^{-3}	2.77×10^{-4}	2.09×10^{-5}	2.78×10^{-4}	3.68×10^{-3}
LeverageSHAP									
Mean	1.38×10^{-3}	3.71×10^{-5}	1.44×10^{-7}	6.32×10^{-3}	1.08×10^{-3}	9.62×10^{-5}	2.83×10^{-5}	1.15×10^{-4}	1.13×10^{-3}
1st Quartile	3.35×10^{-4}	3.07×10^{-7}	7.88×10^{-10}	1.05×10^{-3}	2.74×10^{-4}	5.32×10^{-5}	1.60×10^{-16}	4.41×10^{-5}	2.20×10^{-4}
2nd Quartile	6.62×10^{-4}	2.22×10^{-6}	3.21×10^{-8}	2.73×10^{-3}	1.09×10^{-3}	7.30×10^{-5}	2.70×10^{-6}	9.36×10^{-5}	5.81×10^{-4}
3rd Quartile	1.62×10^{-3}	5.13×10^{-6}	1.29×10^{-7}	7.03×10^{-3}	1.46×10^{-3}	1.08×10^{-4}	2.62×10^{-5}	1.52×10^{-4}	1.30×10^{-3}



Shapley Values: Error vs Sample Complexity
Adult ($n = 12$)
Forest Fires ($n = 13$)
Real Estate ($n = 15$)
Bike Sharing ($n = 16$)
Breast Cancer ($n = 30$)
Independent ($n = 60$)
NHANES ($n = 79$)
Communities ($n = 101$)
ℓ_2-error
Number of Samples (m)
LinearMSR
TreeMSR
KernelSHAP
PermutationSHAP
LeverageSHAP



Probabilistic Values: Error vs Sample Complexity (Regression Forest Ground Truth)
Symmetric Beta
Asymmetric Beta
Weighted Banzhaf
ℓ_2-error
Shapley
B(2,2)
B(4,4)
B(8,8)
B(1,2)
B(1,4)
B(1,8)
Banzhaf
WB(0.6)
WB(0.7)
WB(0.8)
Sample Multiplier (m/n)
LinearMSR
TreeMSR
OFA
WSL
GELS
ARM
WeightedSHAP
LeverageSHAP
KernelSHAP
PermutationSHAP
AME
KernelBanzhaf

Let’s Chat!

rtealwitter@cmc.edu