PDD-SHAP: Fast Approximations for Shapley Values using Functional Decomposition

Arne Gevaert[1 [0000−0003−4130−8151]] and Yvan Saeys[1 [0000−0002−0415−1506]]

Department of Applied Mathematics and Statistics
Ghent University, Ghent, Belgium

Shapley values have gained significant popularity as an explanation method for black-box machine learning models in recent years [11, 7, 6]. However, estimating Shapley values in practice is computationally very expensive. In this work, we exploit properties of the functional ANOVA decomposition [10] to produce a model-agnostic technique for approximating Shapley values with a very low *amortized* cost: if many predictions need to be explained, the cost *per explanation* decreases significantly. We do this by constructing a functional decomposition (ANOVA model), and training it to imitate the black box model being explained. Once the ANOVA model is trained, Shapley values can be estimated orders of magnitude faster than using existing model-agnostic approaches. We empirically show that the cost of training the surrogate model is compensated by the speedup in inference, even for relatively small amounts of explanations.

We denote the black-box model as a function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$. We abbreviate the set $\{1, \ldots, d\}$ to $[d]$, and for a subset $u \subseteq [d]$ we write $-u := [d] \backslash u$. If $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $u \subseteq [d]$, then $\mathbf{z} := \mathbf{x}_u : \mathbf{y}_{-u}$ is defined as $z_j := x_j$ for $j \in u$, and $z_j = y_j$ for $j \notin u$.

Shapley values [11] are a way of fairly distributing a *payout* among participating *players*. Let $\text{val}(u) \in \mathbb{R}$ be the payout of a subset of players $u \subseteq [d]$, with $\text{val}(\emptyset) = 0$. The Shapley value for player $j \in [d]$ is then defined as:

$$\phi_j := \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} (val(u \cup \{j\}) - val(u))$$

The functional ANOVA is a decomposition of the form

$$f(\mathbf{x}) = \sum_{u \subseteq [d]} f_u(\mathbf{x})$$

where each function f_u depends only on the variables $x_j, j \in u$ [3]. Although the original ANOVA decomposition assumes that variables are independent and uniformly distributed, this assumption is not critical. We replace the uniform distribution with a general marginal distribution for each variable: $x_j \sim X_j, \forall j \in [d]$. Note that this approach still assumes independence between variables. We name this variant of the ANOVA decomposition the Partial Dependence Decomposition.

[9] shows that a specific implementation of Shapley values called *Shapley Effects* [12] can be estimated efficiently if an ANOVA decomposition of the black

box function is given. The proof from [9] can be extended to more general implementations of Shapley values and the ANOVA decomposition (including the Partial Dependence decomposition). If we define the value function for an explanation for an input point $\mathbf{x}$ as follows:

$$\text{val}_{\mathbf{x}}(u) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{x}_u : \mathbf{z}_{-u})] - \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{z})]$$

then the corresponding Shapley values can be estimated using the following equation:

$$\phi_j^{\mathbf{x}} = \sum_{\substack{u\subseteq[d] \\ j\in u}} \frac{f_u(\mathbf{x})}{|u|} \tag{1}$$

Our proposed method works by first training a Partial Dependence Decomposition, from which we then compute Shapley values using equation 1. As the number of components f_u that need to be estimated grows exponentially with d, we only estimate the terms f_u for which $|u| < k$, for some $k \in \mathbb{N}$.

Table 1. Runtime (in seconds) for PDD-SHAP vs. subset sampling [13], antithetic sampling [8] and KernelSHAP [6] for 1000 explanations. The rightmost columns correspond to PDD-SHAP for varying values of k. Results are indicated in bold where the sum of training and inference time for PDD-SHAP is lower than the runtime of all 3 alternatives.

Dataset	Subset sampling	Antithetic sampling	Kernel-SHAP	PDD-SHAP (train time + inference time)			
				$k=1$	$k=2$	$k=3$	$k=4$
Adult	103.90	67.44	1060.78	**0.85+0.01**	**6.15+0.02**	**27.33+0.10**	84.29+0.28
Credit	112.16	73.09	1125.11	**1.31+0.01**	**12.42+0.04**	77.96+0.22	370.31+1.06
Superconduct	308.30	97.42	2141.26	**2.98+0.01**	120.09+0.18	N/A	N/A
Housing	68.05	89.25	239.31	**0.43+0.01**	**1.64+0.01**	**4.14+0.02**	**8.03+0.03**
Abalone	108.66	230.72	169.25	**1.08+0.01**	**4.63+0.02**	**11.61+0.03**	**18.80+0.05**

We tested our approach on the adult [5], superconduct [2], UCI German credit [1], California housing [4] and UCI Abalone datasets [1], using a background sample of 100 instances, and using a regression tree to model each f_u[1]. To evaluate our technique, we train a Gradient Boosting model on each dataset and compare the explanations given by PDD-SHAP for 1000 test samples to 3 existing model-agnostic approaches for computing Shapley values: feature subset sampling [13], antithetic sampling [8] and KernelSHAP [6]. These approaches are all implemented in the `shap` package[2].

Table 1 shows the runtime required for each method to generate Shapley values on 1000 instances. We see that the inference time for PDD-SHAP is orders of magnitude lower than the runtime for the existing methods. In many cases even the runtime for training the surrogate model and inference combined is still significantly lower than the runtime for the existing methods.

[1] Implementation available at https://github.com/arnegevaert/pdp-shapley
[2] https://github.com/slundberg/shap

References

1. Dua, D., Graff, C.: UCI machine learning repository (2017)
2. Hamidieh, K.: A data-driven statistical model for predicting the critical temperature of a superconductor. Computational Materials Science **154**, 346–354 (Nov 2018). https://doi.org/10.1016/j.commatsci.2018.07.052
3. Hooker, G.: Discovering additive structure in black box functions. In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04. p. 575. ACM Press, Seattle, WA, USA (2004). https://doi.org/10.1145/1014052.1014122
4. Kelley Pace, R., Barry, R.: Sparse spatial autoregressions. Statistics & Probability Letters **33**(3), 291–297 (May 1997). https://doi.org/10.1016/S0167-7152(96)00140-X
5. Kohavi, R.: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining p. 6 (1996)
6. Lundberg, S., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems **30**, 4766–4775 (2017)
7. Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., Lee, S.I.: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering **2**(10), 749–760 (Oct 2018). https://doi.org/10.1038/s41551-018-0304-0
8. Mitchell, R., Cooper, J., Frank, E., Holmes, G.: Sampling Permutations for Shapley Value Estimation p. 46
9. Owen, A.B.: Sobol' Indices and Shapley Value. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 245–251 (Jan 2014). https://doi.org/10.1137/130936233
10. Roosen, C.B., Friedman, J.H., Owen, A.B.: Visualization And Exploration Of High-Dimensional Functions Using The Functional Anova Decomposition (1995)
11. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
12. Song, E., Nelson, B.L., Staum, J.: Shapley Effects for Global Sensitivity Analysis: Theory and Computation. SIAM/ASA Journal on Uncertainty Quantification **4**(1), 1060–1083 (Jan 2016). https://doi.org/10.1137/15M1048070
13. Štrumbelj, Erik, E., Kononenko, I.: An Efficient Explanation of Individual Classifications using Game Theory. Journal of Machine Learning Research **11**(1), 1–18 (2010)