SUBLINEAR ESTIMATION OF ADDITIVE INTERPRETABILITY UNDER SHAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Additive feature attribution methods explain model predictions as sums of per-feature contributions, but their fidelity is fundamentally limited by how well the model behavior can be represented additively. We formalize this limitation through *additive interpretability*, defined as the best achievable coefficient of determination over the class of additive functions. We reduce the problem to explained-variance estimation and prove that a sublinear learnability estimator can estimate additive interpretability under the Shapley coalition distribution without recovering the optimal attribution vector. For any fixed error tolerance and success probability, the estimator requires $O(\sqrt{d}\log d)$ model evaluations in the high-dimensional regime. Across synthetic and real-world experiments, it accurately estimates additive interpretability with fewer model evaluations than attribution-based baselines. These results enable practitioners to assess whether an additive explanation can faithfully represent a prediction before paying the cost of computing its feature attributions.

1 INTRODUCTION

Additive feature attribution methods are widely used to explain predictions of complex machine learning models by representing the behavior being explained as a sum of per-feature contributions. The resulting explanations assign an importance score to each input feature and are simple to inspect, compare, and communicate. Prominent examples include SHAP (Lundberg & Lee (2017a)) and LIME (Ribeiro et al. (2016)), which construct additive explanations under different assumptions about the behavior being explained and the distribution of perturbed inputs.

Before computing an additive attribution, however, there is a more fundamental question: *can the model's prediction be faithfully explained by an additive function in the first place?* Consider explaining an XOR function at $(1, 1)$. Its prediction depends entirely on the interaction between the two features, which cannot be faithfully represented as independent additive contributions. In such a case, poor explanation fidelity is not necessarily a failure of the attribution method, but a limitation of the additive representation itself.

In this work, we formalize this distinction by defining the **additive interpretability** as the best achievable coefficient of determination $R^2_{\max}(\mathcal{H})$ from a class of additive functions $\mathcal{H}$ in approximating the target model's local behavior. Unlike the fidelity of a particular attribution result, additive interpretability characterizes the representational capacity of the additive explanation class and provides an algorithm-independent upper bound on achievable fidelity. It therefore indicates whether additive feature attribution is an appropriate explanation family for the prediction being explained.

Additive interpretability could be assessed by obtaining and evaluating a near-optimal additive attribution. However, this can require many model evaluations and largely defeats the purpose of using additive interpretability as a pre-explanation diagnostic. We therefore ask whether $R^2_{\max}$ can be estimated more efficiently without finding the optimal attribution.

This problem is closely related to *estimating learnability* in high-dimensional statistics: estimating how well a function class can fit a target without learning the optimal function itself. Kong & Valiant (2018) show that, for linear functions, $R^2_{\max}$ can be estimated with a sublinear number of samples, even when there are too few samples to recover the optimal predictor itself. Their result requires an isotropic (or known-covariance) input distribution, with bounded fourth moments of the inputs, and a bounded joint fourth moment between the inputs and targets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we prove that the Shapley coalition distribution satisfies the structural and moment conditions required for sublinear learnability estimation. Building on this result, we construct an estimator of $R^2_{\max}$ that requires $O(\sqrt{d}\log d)$ model evaluations for fixed accuracy and confidence in the high-dimensional regime, without recovering the optimal attribution vector. Across synthetic and real-world experiments, the estimator accurately estimates additive interpretability in the sublinear regime and requires fewer model evaluations than the evaluated baselines at higher dimensions. Together, these results provide an efficient way to assess the additive representability of local model behavior before computing the corresponding feature attributions.

2 Related Work

Our work connects two largely separate lines of research. The first studies the fidelity and limitations of additive feature explanations, particularly when model behavior contains feature interactions. The second develops estimators of explained variance under linear models.

2.1 Assessing the Limits of Additive Explanations

Additive explanations approximate local model behavior using fixed per-feature contributions (Ribeiro et al. (2016)). This representation cannot fully capture feature interactions, where a feature's effect on the prediction depends on which other features are present. Consequently, using these contributions to infer the model's responses to feature perturbations can be misleading (Bilodeau et al. (2024)).

One approach to assessing this limitation is to evaluate a computed explanation. LEAF by Amparore et al. (2021) and REVEL by Sevillano-García et al. (2023) assess agreement between a fitted local explanation and the model over a sampled neighborhood, using classification agreement and predicted probability agreement. Other metrics evaluate feature importance rankings through prediction changes as highly ranked input regions are removed or restored (Samek et al. (2017); Petsiuk et al. (2018)). However, observed error reflects both representational limitations and shortcomings of the returned explanation. Evaluation also requires computing the explanation first, incurring its cost before establishing whether an additive representation is appropriate.

A complementary approach explicitly characterizes feature interactions. SHAP interaction values quantify pairwise contributions to individual predictions (Lundberg & Lee (2017b)). Subsequent work supports higher-order Shapley interactions (Muschalik et al. (2024)) and directional feature relationships (Masoomi et al. (2022)). These methods reveal dependencies omitted by individual feature contributions, but estimating many interactions can require substantial computation.

More directly related to our objective, functional ANOVA and related methods measure departure from additive structure. Hooker (2004) quantifies the loss from projection onto an additive model class. Such analyses typically characterize the function over a reference input distribution, whereas our focus is the local behavior associated with explaining an individual prediction. We define additive interpretability and estimate it using coalitions sampled from the Shapley distribution, requiring a sublinear number of model evaluations without recovering the optimal additive explanation.

2.2 Estimating Unexplained Variance under Linear Models

Estimating explained and unexplained variance is a classical statistical problem. Under a correctly specified linear model, ordinary least squares (OLS) estimates the noise variance from fitted residuals, with a correction for the number of estimated coefficients. This approach requires positive residual degrees of freedom and its unbiasedness relies on a linear conditional mean and constant noise variance. Other methods avoid coefficient recovery: Dicker (2014) develops moment-based estimators with sublinear sample requirements, while EigenPrism constructs variance estimates and confidence intervals (Janson et al. (2017)). Their core guarantees nevertheless assume Gaussian linear models, which do not generally describe nonlinear coalition responses.

Sparse high-dimensional methods address this regime through regularized estimation and bias correction. For example, CHIVE estimates explained variance using a calibrated estimator and can incorporate additional unlabelled observations (Cai & Guo (2020)). Its guarantees rely on a sparse regression vector and suitable design and noise conditions, which need not hold for the optimal additive explanation.

More generally, Kong & Valiant (2018) permit nonlinear targets and estimate the minimum linear prediction error without recovering the coefficients or assuming sparsity. Their guarantee depends on known covariance and controlled fourth moments. We establish the relevant covariance and input moment properties for the coalition distribution and, under explicit target moment conditions, obtain a sublinear estimation guarantee for additive interpretability.

3 Additive Interpretability

We consider the problem of explaining the prediction of a model f at an input x using an additive function defined over a local perturbation space. Let $z \in \{0,1\}^d$ denote a coalition indicating which features of x are present, and let $h_x(z)$ map the coalition to the corresponding perturbed input. The model behavior to be explained is therefore

$$y(z) = f(h_x(z)).$$

An explanation setting additionally specifies a distribution $p(z)$ over coalitions, which determines the regions of the perturbation space over which approximation fidelity is evaluated. Throughout this work, all quantities are defined with respect to a fixed model f, input x, perturbation mapping h_x, and coalition distribution $p(z)$.

An additive explanation represents this behavior using a function of the form

$$g_\theta(z) = \theta_0 + \sum_{i=1}^{d} \theta_i z_i,$$

where θ_i denotes the contribution associated with feature i. Let

$$\mathcal{H} = \left\{ g_\theta : g_\theta(z) = \theta_0 + \theta^\top z \right\}$$

denote the corresponding class of additive functions. For a particular additive explanation $g \in \mathcal{H}$, we measure its fidelity using the coefficient of determination

$$R^2(g) = 1 - \frac{\mathbb{E}_{z\sim p}\left[\left(y(z) - g(z)\right)^2\right]}{\mathrm{Var}_{z\sim p}[y(z)]}.$$

This evaluates the fidelity of a particular additive explanation. However, poor observed fidelity may arise either because the model behavior is intrinsically non-additive or because the attribution method fails to recover the best additive approximation. To separate these two factors, we consider the best fidelity achievable by the entire additive function class.

Definition 1 (Additive Interpretability)**.** *Given a model f, an input x, a perturbation mapping h_x, a coalition distribution $p(z)$, and an additive function class $\mathcal{H}$, the **additive interpretability** of the prediction is the best achievable coefficient of determination within $\mathcal{H}$:*

$$\mathrm{AI}(f, x; \mathcal{H}, p) = R^2_{\max}(\mathcal{H}) = \sup_{g\in\mathcal{H}} R^2(g).$$

Additive interpretability therefore characterizes the representational capacity of the additive explanation class rather than the quality of a particular attribution estimation algorithm. Thus, $R^2_{\max}$ provides an algorithm-independent upper bound on the fidelity achievable by additive explanations. A low value indicates that the model behavior cannot be faithfully represented additively, regardless of how accurately the attribution coefficients are estimated. A high value indicates that a high-fidelity additive representation exists, although a particular attribution method may still fail to recover it.

4 Estimating Additive Interpretability

This section investigates whether $R^2_{\max}$ can be estimated without recovering the optimal additive explanation. We first reduce additive interpretability to estimating the variance explained by the optimal additive approximation. We then adopt the sublinear explained-variance estimator of Kong & Valiant (2018), and prove that the Shapley coalition distribution satisfies its key moment condition with only logarithmic dependence on the dimensionality. This yields our main result: an explicit error bound for estimating additive interpretability from a sublinear number of model evaluations.

4.1 Reducing Additive Interpretability to Explained Variance

We first identify the scalar quantity that must be estimated to obtain additive interpretability. Recall that $y(z) = f(h_x(z))$. Define

$$\tilde{z} = z - \mathbb{E}_p[z], \qquad \tilde{y} = y(z) - \mathbb{E}_p[y(z)],$$

and let

$$\Sigma = \mathbb{E}_p[\tilde{z}\tilde{z}^\top], \qquad \mu = \mathbb{E}_p[\tilde{z}\tilde{y}].$$

Because the additive function class $\mathcal{H}$ includes an intercept, optimizing over $g_\theta \in \mathcal{H}$ is equivalent to fitting a linear function between $\tilde{z}$ and $\tilde{y}$. Assuming Σ is invertible, the population least-squares coefficients are

$$\theta^\star = \Sigma^{-1}\mu.$$

The corresponding minimum squared error is

$$\min_\theta \mathbb{E}_p\left[(\tilde{y} - \theta^\top \tilde{z})^2\right] = \mathrm{Var}_p[y(z)] - \mu^\top \Sigma^{-1}\mu.$$

Therefore,

$$R^2_{\max} = \frac{\mu^\top \Sigma^{-1}\mu}{\mathrm{Var}_p[y(z)]}.$$

We define

$$S := \mu^\top \Sigma^{-1}\mu$$

as the variance explained by the optimal additive approximation. Estimating additive interpretability thus reduces to estimating S together with the total variance $\mathrm{Var}_p[y(z)]$. Since the latter is a standard scalar variance, the main challenge is estimating S without recovering the full coefficient vector $\theta^\star$.

4.2 Estimating the Explained Variance

We next introduce the explained-variance estimator of Kong & Valiant (2018). Their estimator directly estimates the scalar S through pairwise products of independent samples, avoiding estimation of the underlying d-dimensional linear predictor.

Define the whitened coalition

$$q = \Sigma^{-1/2}\tilde{z}.$$

Then

$$\mathbb{E}_p[q] = 0, \qquad \mathbb{E}_p[qq^\top] = I, \qquad \text{and} \qquad S = \left\|\mathbb{E}_p[q\tilde{y}]\right\|_2^2.$$

For two independent samples $(q_i, \tilde{y}_i)$ and $(q_j, \tilde{y}_j)$,

$$\mathbb{E}\left[\tilde{y}_i\tilde{y}_j q_i^\top q_j\right] = \mathbb{E}[\tilde{y}_i q_i]^\top \mathbb{E}[\tilde{y}_j q_j] = S.$$

This motivates the pairwise estimator

$$\widehat{S} = \frac{1}{\binom{n}{2}} \sum_{i<j} \tilde{y}_i\tilde{y}_j q_i^\top q_j.$$

Together with an estimate $\widehat{V}_y$ of $\mathrm{Var}_p[y(z)]$, we estimate additive interpretability as

$$\widehat{R}^2_{\max} = \frac{\widehat{S}}{\widehat{V}_y}.$$

The following proposition summarizes the relevant result of Kong and Valiant in our notation.

Proposition 1 (Theorem 4 in Kong & Valiant (2018))**.** *The estimator $\widehat{S}$ is unbiased:*

$$\mathbb{E}[\widehat{S}] = S.$$

Let

$$r = \frac{\tilde{y}}{\sqrt{\mathrm{Var}_p[y(z)]}}.$$

Suppose there exists $C \geq 1$ such that, for all $u, v \in \mathbb{R}^d$,

$$\mathbb{E}\left[(q^\top u)^2 (q^\top v)^2\right] \leq C\mathbb{E}[(q^\top u)^2]\mathbb{E}[(q^\top v)^2],$$

and

$$\mathbb{E}\left[r^2 (q^\top u)^2\right] \leq C\mathbb{E}[(q^\top u)^2]\mathbb{E}[r^2].$$

Then, with probability at least $1-\tau$, the estimation error is bounded by

$$O\left(C\frac{\sqrt{d+n}}{\tau n}\right).$$

The covariance of the Shapley coalition distribution is known analytically, so centering and whitening do not require estimating Σ by sampling. The remaining nontrivial requirement in Proposition 1 is the fourth-moment condition of the whitened coalition distribution. We therefore next determine how the corresponding constant C scales under Shapley sampling.

4.3 Fourth-Moment Bound under Shapley Sampling

We now establish the key Shapley-specific result needed to instantiate Proposition 1. Although Shapley sampling places relatively high probability on coalitions close to the empty and full sets, we show that whitening increases the relevant fourth moments by only a logarithmic factor.

As in KernelSHAP, the empty and full coalitions are handled separately. Let z denote a nontrivial coalition sampled from the normalized Shapley distribution. By symmetry,

$$\mathbb{E}_p[z] = \frac{1}{2}\mathbf{1}.$$

We use the centered representation

$$z_c = 2z - \mathbf{1}.$$

Its covariance has the permutation-symmetric form

$$\Sigma_c = \alpha I + \beta \mathbf{1}\mathbf{1}^\top, \qquad \text{where} \qquad \alpha = \frac{2}{H_{d-1}},$$

and H_{d-1} is the $(d-1)$-th harmonic number. The derivation is provided in Appendix A.3.

Let

$$q = \Sigma_c^{-1/2} z_c.$$

By Cauchy–Schwarz,

$$\mathbb{E}\left[(q^\top u)^2 (q^\top v)^2\right] \leq \sqrt{\mathbb{E}[(q^\top u)^4]\mathbb{E}[(q^\top v)^4]}.$$

Thus, it suffices to bound the maximum fourth moment of a one-dimensional projection of q.

The proof proceeds by undoing the whitening and conditioning on the coalition size $s = |z|$. Conditional on s, the coalition is a uniformly sampled subset without replacement. This permits the fourth moment to be expressed through low-order inclusion probabilities. Averaging these terms over the Shapley distribution of coalition sizes introduces a factor of order $1/H_{d-1}$, while the whitening constraint contributes a factor of order H_{d-1}^2. Combining these terms gives an overall $O(H_{d-1})$ fourth-moment bound. The complete derivation is provided in Appendix A.

Lemma 1 (Fourth-moment bound under Shapley sampling)**.** *For the centered and whitened Shapley coalition distribution,*

$$\sup_{\|u\|_2=1} \mathbb{E}_p\left[(q^\top u)^4\right] = O(H_{d-1}) = O(\log d).$$

Consequently, the fourth-moment condition in Proposition 1 is satisfied with

$$C = O(\log d).$$

We additionally assume that the standardized model output r has bounded fourth moment,

$$\mathbb{E}_p[r^4] \le \kappa,$$

for some constant $\kappa = O(1)$.

By the Cauchy–Schwarz inequality and Lemma 1, we obtain

$$\mathbb{E}_p\left[r^2(q^\top u)^2\right] \le \sqrt{\mathbb{E}_p[r^4]\,\mathbb{E}_p[(q^\top u)^4]} \qquad \text{and} \qquad \mathbb{E}_p\left[r^2(q^\top u)^2\right] = O(\sqrt{\log d})\|u\|_2^2.$$

Because q is whitened and r is standardized,

$$\mathbb{E}_p[(q^\top u)^2] = \|u\|_2^2, \qquad \mathbb{E}_p[r^2] = 1.$$

Therefore,

$$\mathbb{E}_p\left[r^2(q^\top u)^2\right] \le O(\sqrt{\log d})\,\mathbb{E}_p[(q^\top u)^2]\mathbb{E}_p[r^2].$$

Hence the joint moment condition in Proposition 1 is satisfied with $C = O(\sqrt{\log d})$. Together with the input fourth-moment bound $C = O(\log d)$ from Lemma 1, both conditions of Proposition 1 hold with the common bound

$$C = O(\log d).$$

Lemma 1 is the key technical step that allows the sublinear estimator of Kong and Valiant to remain effective under the highly non-uniform Shapley coalition distribution.

4.4 Estimation Error and Sample Complexity

We now combine Proposition 1 with the Shapley-specific moment bound in Lemma 1. Because additive interpretability is the ratio

$$R^2_{\max} = \frac{S}{V_y}, \qquad V_y := \mathrm{Var}_p[y(z)],$$

we must account for estimation of both S and V_y.

Under the bounded-fourth-moment assumption

$$r = \frac{y - \mathbb{E}_p[y]}{\sqrt{V_y}}, \qquad \mathbb{E}_p[r^4] \le \kappa,$$

the unbiased sample variance satisfies, with probability at least $1 - \tau$,

$$\left|\frac{\widehat{V}_y}{V_y} - 1\right| = O\left(\sqrt{\frac{\kappa}{n\tau}}\right).$$

The proof is given in Appendix A.9. Combining this result with Proposition 1 yields the following Shapley-specific guarantee.

Theorem 1 (Estimation error under Shapley sampling)**.** *Let the coalitions be sampled independently from the normalized Shapley distribution, and suppose the moment conditions above hold with $\mathbb{E}_p[r^4] \le \kappa$. Then, with probability at least $1 - \tau$,*

$$\left|\widehat{R}^2_{\max} - R^2_{\max}\right| = O\left(\log d\,\frac{\sqrt{d+n}}{\tau n} + \sqrt{\frac{\kappa}{n\tau}}\right).$$

Solving the bound in Theorem 1 for a desired additive error ϵ gives the following sample complexity.

Corollary 1 (Sample complexity)**.** *To estimate additive interpretability to additive error ϵ with probability at least $1 - \tau$, it suffices to use*

$$n = O\left(\frac{\sqrt{d}\log d}{\tau\epsilon} + \frac{\log^2 d}{\tau^2\epsilon^2} + \frac{\kappa}{\tau\epsilon^2}\right)$$

model evaluations. When $\kappa = O(1)$, the final term is absorbed by the second term, yielding

$$n = O\left(\frac{\sqrt{d}\log d}{\tau\epsilon} + \frac{\log^2 d}{\tau^2\epsilon^2}\right).$$

For fixed accuracy and confidence, the first term dominates in the high-dimensional regime, yielding $O(\sqrt{d}\log d)$ model evaluations. For smaller d or more stringent accuracy requirements, the second term can dominate. Thus, additive interpretability can be estimated in the high-dimensional regime using a sublinear number of model evaluations, without recovering the optimal attribution vector.

5 EXPERIMENTS

We empirically evaluate three aspects of the proposed estimator. First, we examine whether its estimation error decreases in the sublinear regime, where recovering the full attribution vector is generally difficult. Second, we compare its empirical sample requirement with attribution-based, regression-based, and existing explained-variance estimators on real and synthetic data. Third, we study how its sample requirement changes with the dimension on a controlled family of synthetic problems. These experiments evaluate the accuracy and sample efficiency of the estimator; they do not directly evaluate downstream decisions about whether an additive explanation should be used.

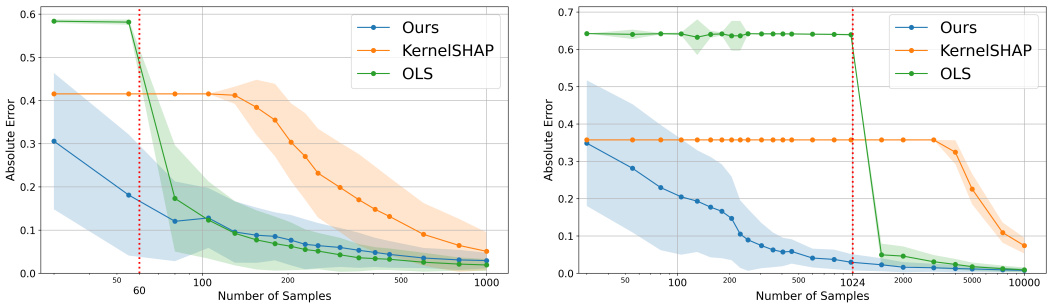
We compare against four baselines. KernelSHAP first estimates an attribution vector and then evaluates its fidelity on held-out coalitions using an 80/20 training-test split. Ordinary Least Squares (OLS) estimates unexplained variance from fitted residuals when $n > d$; for $n \leq d$, we use a regularized regression plug-in estimate. We also include CHIVE (Cai & Guo (2020)) and EigenPrism (Janson et al. (2017)) as representatives of high-dimensional explained-variance estimators. Their theoretical guarantees rely, respectively, on sparsity and Gaussian linear-model assumptions that need not hold for coalition responses. Their results should therefore be interpreted as empirical reference points under model misspecification rather than comparisons within their guaranteed settings. Appendix B.5 introduces the details of the estimator implementation.

We evaluate the estimators using models trained on six real-world datasets and a collection of synthetic functions, combining realistic prediction tasks with settings in which the target quantities can be computed directly. The real-world datasets are obtained from the UCI Machine Learning Repository. For each dataset, we train a neural network with three hidden layers and select 10 target predictions for explanation. For each prediction, we construct a high-budget reference estimate of $R^2_{\max}$ using KernelSHAP with an independent set of sampled coalitions. We increase the reference budget until the estimate satisfies the convergence criterion. The dataset, model, and estimation details are described in Appendix B.2. The synthetic functions are multilinear polynomials constructed so that both their SHAP attributions and $R^2_{\max}$ can be computed directly from the polynomial coefficients.

For a fixed prediction, estimator, and random seed, we use binary search over budgets between 10 and $s_{\max} = 10{,}000$ to approximate the smallest tested budget at which the absolute error is at most in different threshold $\epsilon \in \{0.1, 0.05, 0.025\}$. We repeat this procedure for 10 independent seeds and report the eighth order statistic of the resulting budgets. Thus, the reported value approximates the budget by which $1 - \tau = 80\%$ observed runs have reached the target error. Because a realized error trajectory need not be monotone in the sample size, this procedure is an empirical first-passage approximation rather than an estimate of the smallest budget that guarantees the target accuracy. We average this quantity across 10 predictions or synthetic functions and report its mean and standard deviation. We report the results with the tightest threshold $\epsilon = 0.025$ in this section and add the full results in Appendix C.

5.1 CONVERGENCE BEHAVIOUR

Figure 1 shows the absolute estimation error as the sample budget increases for one prediction from each of two datasets. Curves are averaged over 50 independent runs. The proposed estimator improves substantially while $n < d$ and approaches the reference value before $n = d$ in both examples. In contrast, the attribution-first KernelSHAP baseline and the regression plug-in estimator improve more slowly in this regime. After n exceeds d, the regression-based estimator improves rapidly. These examples illustrate the regime targeted by our method: estimating the scalar $R^2_{\max}$ when the available budget is insufficient to stably recover a d-dimensional attribution vector.



(a) Sonar (d = 60)

(b) Qsar Toxic (d = 1024)

Figure 1: Convergence comparison on two datasets (shown in logarithmic scale). The x-axis shows increasing number of samples, while the y-axis shows the absolute error of the estimator. The line shows the mean error across 50 runs, while the shaded area shows the standard deviation. The dashed red line shows the dimensionality of the dataset.

Table 1: Empirical model-evaluation budget required to estimate $R^2_{\max}$ with absolute error at most 0.025 in at least 8 of 10 independent runs, corresponding to an observed failure rate of $\tau = 0.2$.

Dataset	d	Ours	KernelSHAP	OLS	CHIVE	EigenPrism
Australian	14	561 ± 154	595 ± 555	$\mathbf{254 \pm 120}$	650 ± 474	587 ± 379
Spectfheart	44	600 ± 526	1941 ± 1715	$\mathbf{555 \pm 539}$	3312 ± 1942	648 ± 293
Sonar	60	611 ± 323	2102 ± 1033	$\mathbf{443 \pm 341}$	4668 ± 1390	739 ± 348
Tuandromd	240	$\mathbf{629 \pm 487}$	4846 ± 1266	3525 ± 2454	1708 ± 3163	5251 ± 1867
Parkinsons	753	$\mathbf{1539 \pm 465}$	$> s_{max}$	5952 ± 2747	8027 ± 2275	5255 ± 3948
Qsar Toxic	1024	$\mathbf{992 \pm 377}$	$> s_{max}$	3407 ± 3019	3215 ± 4442	$> s_{max}$
Synthetic	50	422 ± 122	878 ± 506	$\mathbf{160 \pm 82}$	896 ± 685	524 ± 273
Synthetic	100	394 ± 150	1488 ± 482	$\mathbf{362 \pm 105}$	2990 ± 1079	887 ± 473
Synthetic	250	$\mathbf{616 \pm 148}$	5624 ± 2694	1861 ± 809	8421 ± 1145	2278 ± 1447
Synthetic	500	$\mathbf{869 \pm 641}$	7431 ± 2676	4596 ± 3012	6060 ± 4824	7425 ± 3499
Synthetic	1000	$\mathbf{1411 \pm 1470}$	9874 ± 250	6313 ± 2302	4098 ± 4819	9148 ± 1703
Synthetic	2000	$\mathbf{1560 \pm 794}$	$> s_{max}$	9933 ± 132	8020 ± 3958	8087 ± 3824

5.2 Comparison among other estimators

Table 1 shows that on the low-dimensional real datasets, OLS or the regression plug-in baseline requires the fewest samples. The proposed estimator becomes more favorable as the dimension increases: it has the lowest reported mean budget on Tuandromd, Parkinsons, and Qsar Toxic. On the synthetic benchmark, the same transition occurs between $d = 100$ and $d = 250$, after which the proposed estimator has the lowest reported mean at every tested dimension.

CHIVE and EigenPrism generally require larger budgets in the higher-dimensional settings. This behavior is consistent with the mismatch between their assumptions and the coalition-response setting, although these experiments do not isolate which assumption is responsible for the observed difference. Overall, the results support the intended division of regimes: conventional regression estimators remain competitive at small d, whereas direct estimation of $R^2_{\max}$ is advantageous when d is large relative to the evaluation budget.

5.3 Empirical scaling with dimension

Figure 2 shows the empirical budget required to reach each error threshold as the dimension increases from 100 to 5,000. Fitting the empirical relationship $s(d) = cd^e$ gives $e = 0.40$, with a 95% confidence interval of $[0.32, 0.47]$. Among the candidate curves, $d^{0.4}$, $\sqrt{d}$, and $\sqrt{d}\log d$ obtain RMSEs of 170, 177, and 186, respectively, compared with 231 for the linear fit. Over the tested

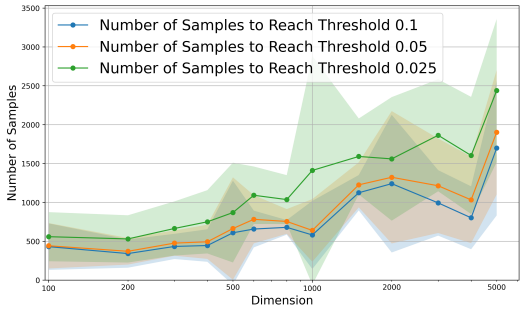


Figure 2: Empirical model-evaluation budget required to attain each absolute-error threshold as dimension increases on the synthetic benchmark (shown in logarithmic scale). The curves show that the required budget grows substantially slower than a linear function $n = O(d)$.

range, the observed sample requirement therefore grows more slowly than the linear reference and is compatible with the derived sublinear upper bound. The differences among the sublinear fits are not sufficient to identify a particular asymptotic rate.

6 Discussion

Additive interpretability characterizes representational fidelity, but its practical significance depends on the intended use of an explanation. Under the chosen perturbation mapping and coalition distribution, a low $R^2_{\max}$ means that every additive explanation leaves a substantial fraction of output variance unexplained. The absolute approximation error may nevertheless be small when this variance is small. Conversely, a high value indicates that a faithful additive approximation exists, without guaranteeing that a particular attribution method recovers it. An important next step is to investigate whether, and under what conditions, this measure predicts explanation usefulness in tasks such as simulating model predictions or deciding when to rely on them.

The prevalence and causes of low additive interpretability also remain open questions. Studying its distribution across architectures, datasets, and individual predictions would clarify how widespread substantial representational limitations are and which properties of local model behavior account for them. Such comparisons should also examine masking choices, since different masking methods can produce different additive interpretability values for the same model and prediction.

A further limitation concerns estimation: our analysis assumes that the standardized model output has bounded fourth moment. This moment–the output kurtosis–can be large in extreme cases, weakening the guarantee and potentially increasing sample requirements, although we did not encounter such cases in our experiments. In practice, assessing output variance is a useful preliminary check: additive interpretability is undefined at zero variance, while very small variance warrants care in normalization and estimation. These considerations motivate reporting additive interpretability alongside output variance and estimation uncertainty to support its use in practice.

7 Conclusion

We introduced additive interpretability to distinguish representational limitations from attribution estimation errors. We proved that the estimator of Kong & Valiant (2018) could be used to assess the best achievable additive fidelity under Shapley sampling using $O(\sqrt{d}\log d)$ model evaluations for fixed accuracy and confidence, without recovering feature attributions. Experiments on synthetic and real-world data support this scaling and demonstrate lower sample requirements than the evaluated baselines at higher dimensions. These results enable assessing whether local model behavior admits a faithful additive explanation before computing the explanation itself.

Reproducibility Statement

The supplement to this submission contains all the code and datasets that were used to produce the shown results. To reproduce the results shown, follow the README in the submission code folder. All the experiments can be run from the main.py file.

AI Usage Declaration

In this work, we used generative AI tools, including Claude, ChatGPT/Codex and GitHub Copilot, to help formulate and check mathematical arguments, suggest intermediate proof steps, provide feedback on the experimental methodology, assist with code implementation and debugging, and refine the interpretation and presentation of results. We also used these tools for literature discovery, manuscript drafting and editing, and formatting equations and tables. We did not use generative AI to generate synthetic datasets, propose hypotheses, translate text, clean datasets, or conduct qualitative data analysis. The authors checked all cited sources, independently verified every mathematical claim, reviewed and tested all AI-assisted code, and independently conducted and evaluated the experiments. We take responsibility for the final content of this work, including all AI-assisted text, claims, code, and artifacts.

References

Elvio Amparore, Alan Perotti, and Paolo Bajardi. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Computer Science*, 7:e479, April 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.479. URL http://dx.doi.org/10.7717/peerj-cs.479.

Davide Ballabio, Francesca Grisoni, Viviana Consonni, and Roberto Todeschini. Integrated QSAR models to predict acute oral systemic toxicity. *Molecular Informatics*, 38(8–9):1800124, 2019. doi: 10.1002/minf.201800124.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024. doi: 10.1073/pnas.2304406120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2304406120.

Parthajit Borah, Dhruba K. Bhattacharyya, and Jugal K. Kalita. Malware dataset generation and evaluation. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pp. 1–6, 2020. doi: 10.1109/CICT51604.2020.9312053.

T. Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):391–419, 2020. doi: 10.1111/rssb.12357.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3457–3465. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/covert21a.html.

Lee H. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014. doi: 10.1093/biomet/ast065.

R. Paul Gorman and Terrence J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1(1):75–89, 1988. doi: 10.1016/0893-6080(88)90023-8.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 575–580, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. doi: 10.1145/1014052.1014122. URL https://doi.org/10.1145/1014052.1014122.

Lucas Janson, Rina Foygel Barber, and Emmanuel J. Candès. Eigenprism: Inference for high-dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1037–1065, 2017. doi: 10.1111/rssb.12203.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *ICLR 2015*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020. URL https://arxiv.org/abs/2009.07896.

Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8bd39eae38511daad6152e84545e504d-Paper.pdf.

Lukasz A. Kurgan, Krzysztof J. Cios, Ryszard Tadeusiewicz, Marek Ogiela, and Lucy S. Goodenday. Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial Intelligence in Medicine*, 23(2):149–169, 2001. doi: 10.1016/S0933-3657(01)00082-3.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017a. Curran Associates Inc. ISBN 9781510860964. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

Scott M. Lundberg and Su-In Lee. Consistent feature attribution for tree ensembles. *CoRR*, abs/1706.06060, 2017b. URL http://arxiv.org/abs/1706.06060.

Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P. Hersh, Edwin K. Silverman, Peter J. Castaldi, Stratis Ioannidis, and Jennifer Dy. Explanations of black-box models based on directional feature interactions. In *10th International Conference on Learning Representations (ICLR)*, 2022. URL https://iclr.cc/virtual/2022/poster/6731.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: shapley interactions for machine learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. URL https://dl.acm.org/doi/10.5555/3737916.3742057.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. *ArXiv*, abs/1806.07421, 2018. URL https://api.semanticscholar.org/CorpusID:49324724.

Ross Quinlan. Statlog (australian credit approval). UCI Machine Learning Repository, 1987.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

C. Okan Sakar, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tutuncu, Tarkan Aydin, M. Erdem Isenkul, and Hulya Apaydin. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, 74:255–263, 2019. doi: 10.1016/j.asoc.2018.10.022.

Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2660–2673, 11 2017. doi: 10.1109/TNNLS.2016.2599820.

Iván Sevillano-García, Julián Luengo, Francisco Herrera, and Alexander Hošovský. Revel framework to measure local linear explanations for black-box models: Deep learning image classification case study. *Int. J. Intell. Syst.*, 2023, January 2023. ISSN 0884-8173. doi: 10.1155/2023/8068569. URL https://doi.org/10.1155/2023/8068569.

A BOUNDING C UNDER THE SHAPLEY DISTRIBUTION

To show the applicability of the estimator derived by Kong & Valiant (2018) to our setting, it remains to show that we can bound the value of C in the input inequality of Proposition 1. The proof will proceed as follows:

We first describe the variable transformation from the KernelSHAP input masks to the input used in the estimator. We then show that the inequality can be reduced to finding an upper bound on the fourth moment of the inner product of the input variable and a normed vector. We then move the whitening matrix to the normalization condition. By looking at the covariance structure, we then derive useful bounds on the conditioned vector. Finally, we decompose the inner product into two terms and bound the expected value of each.

A.1 VARIABLE TRANSFORMATION

We recall that the input variables x to which we apply the bound are the centered and whitened shapley coalitions z. Centering is achieved by the affine transform $z_c = 2z - \mathbf{1}$. The covariance matrix of z_c is then equal to $4\,\mathrm{Cov}(z)$. Thus writing $\Sigma = 4\,\mathrm{Cov}(z)$ we get $x = \Sigma^{-1/2} z_c$.

A.2 REDUCTION TO A SINGLE FOURTH MOMENT

Next, we simplify the condition on C. By the Cauchy-Schwarz inequality we get:

$$\mathbb{E}\big[(x^\top u)^2 (x^\top w)^2\big] \leq \sqrt{\mathbb{E}\left[(x^\top u)^4\right] \mathbb{E}\left[(x^\top w)^4\right]} \tag{1}$$

Moreover, since x is whitened we get $\mathbb{E}[(x^\top u)^2] = \|u\|_2^2$. Putting this together we get:

$$\sqrt{\mathbb{E}\left[(x^\top u)^4\right] \mathbb{E}\left[(x^\top w)^4\right]} \leq C \,\|u\|_2^2 \,\|w\|_2^2 \tag{2}$$

$$\Rightarrow \mathbb{E}\big[(x^\top u)^2 (x^\top w)^2\big] \leq C\, \mathbb{E}[(x^\top u)^2]\, \mathbb{E}[(x^\top w)^2] \tag{3}$$

$$\Rightarrow C \leq \sup_{\|u\|_2 = 1} \mathbb{E}\left[(x^\top u)^4\right] \tag{4}$$

In other words, if we can find a C that satisfies the first line then it also satisfies the inequality and since we can divide by the norm of the vector u and choose $u = w$ for the worst case, we get that $C = \sup_{\|u\|_2 = 1} \mathbb{E}\left[(x^\top u)^4\right]$ satisfies the inequality. Thus, the minimal C must satisfy $C \leq \sup_{\|u\|_2 = 1} \mathbb{E}\left[(x^\top u)^4\right]$. Putting this together with the transformation described above we get:

$$C \leq \sup_{\|u\|_2 = 1} \mathbb{E}\left[(x^\top u)^4\right] \tag{5}$$

$$\Leftrightarrow C \leq \sup_{\|u\|_2 = 1} \mathbb{E}\left[((\Sigma^{-1/2} z_c)^\top u)^4\right] \tag{6}$$

$$\Leftrightarrow C \leq \sup_{v^\top \Sigma v = 1} \mathbb{E}\left[(z_c^\top v)^4\right] \tag{7}$$

A.3 COVARIANCE STRUCTURE

We now delve deeper into the exact values of Σ and the constraint on v that can thus be derived from $v^\top \Sigma v = 1$. We first recall that since the shapley distribution is over the coalitions size $s = |z|$ it is symmetric over all dimensions and in particular we can write $\Sigma = \alpha I + \beta \mathbf{1}\mathbf{1}^\top$.

Writing $b := \mathbf{1}^\top v$, the constraint $v^\top \Sigma v = 1$ can then be written as $\alpha \|v\|_2^2 + \beta b^2 = 1$, which implies

$$\|v\|_2^2 \leq \frac{1}{\alpha} \qquad \text{and} \qquad b^2 \leq \frac{1}{\beta}\,. \tag{8}$$

The values of α and β can be computed as follows:

$$\Sigma = 4\,\mathrm{Cov(Z)} = 4(\mathbb{E}\left[ZZ^\top\right] - \mathbb{E}[Z]\mathbb{E}[Z]^\top) = 4\mathbb{E}\left[ZZ^\top\right] - \mathbf{1}\mathbf{1}^\top$$

The second moment matrix $A := \mathbb{E}\left[ZZ^\top\right]$ has been shown to have diagonal $A_{ii} = \frac{1}{2}$ and off-diagonal $A_{ij} = \frac{1}{d(d-1)}\,N/Z$ with $N = \sum_{k=2}^{d-1} \frac{k-1}{d-k}$ and $Z = \sum_{s=1}^{d-1} \frac{1}{s(d-s)}$ (Covert & Lee, 2021, Supp., App. A). Letting $H_d := \sum_{k=1}^{d} \frac{1}{k}$ be the d-th harmonic number partial fractions give

$$Z = \frac{1}{d}\sum_{k=1}^{d-1}\left(\tfrac{1}{k} + \tfrac{1}{d-k}\right) = \frac{2H_{d-1}}{d}, \tag{9}$$

and the substitution $j = k-1$, $\frac{j}{d-1-j} = \frac{d-1}{d-1-j} - 1$, yields $N = (d-1)H_{d-2} - (d-2)$. Thus

$$A_{ij} = \frac{(d-1)H_{d-2} - (d-2)}{2(d-1)H_{d-1}} = \frac{1}{2} - \frac{1}{2H_{d-1}},$$

using $H_{d-2} = H_{d-1} - \frac{1}{d-1}$.

Writing $A = (\frac{1}{2} - A_{ij})I + A_{ij}\mathbf{1}\mathbf{1}^\top$, we get $\Sigma = (2 - 4A_{ij})I + (4A_{ij} - 1)\mathbf{1}\mathbf{1}^\top$. Thus, we see that

$$\alpha = 2 - 4A_{ij} = \tfrac{2}{H_{d-1}} \qquad \text{and} \qquad \beta = 1 - \alpha.$$

As $\beta = 1 - \frac{2}{H_{d-1}} \geq \frac{1}{4}$ for $d \geq 11$ we see that $b^2 \leq \frac{1}{\beta} \leq 4 \leq O(1)$.

A.4 Decomposition of the Inner Product

Let S be the coalition associated with z, $s = |z|$ (i.e. the coalition size), $\bar{v}_i = v_i - \frac{b}{d}$ (so $\sum_i \bar{v}_i = 0$ and $\|\bar{v}\|_2^2 \leq \|v\|_2^2 \leq \alpha^{-1}$), $R = 2s - d$, and $\bar{Y} = \sum_{i \in S} \bar{v}_i$. Then

$$z_c^\top v = 2\sum_{i \in S} v_i - b = 2\bar{Y} + \tfrac{R}{d}\,b. \tag{10}$$

By convexity of $t \mapsto t^4$ and Jensen's inequality, $(\frac{P+Q}{2})^4 \leq \frac{(P^4+Q^4)}{2}$ which gives $(P+Q)^4 \leq 8(P^4+Q^4)$. Thus, with $P = 2\bar{Y}$, $Q = \frac{R}{d}b$ and the linearity of $\mathbb{E}$,

$$\mathbb{E}[(z_c^\top v)^4] \leq 128\,\mathbb{E}[\bar{Y}^4] + \frac{8b^4}{d^4}\mathbb{E}[R^4]. \tag{11}$$

Since b^2 is in $O(1)$, and $R^4 \leq d^4$, the right-hand term is at most $O(1)$. It remains to bound $\mathbb{E}[\bar{Y}^4]$.

A.5 Bounding the Expected Value

Write $s^{\underline{r}} = s(s-1)\cdots(s-r+1)$ and $p_r(s) = s^{\underline{r}}/d^{\underline{r}} = P(i_1, \ldots, i_r \in S)$ for distinct indices, the inclusion probability of an r-set under uniform s-subset sampling. For visual clarity we shorten $p_r(s)$ to p_r . Let

$$V_2 = \|\bar{v}\|_2^2 = \sum_i \bar{v}_i^2, \qquad V_4 = \|\bar{v}\|_4^4 = \sum_i \bar{v}_i^4.$$

Lemma 2 (Exact SWoR fourth moment)**.** *For a centred population ($\sum_i \bar{v}_i = 0$) and uniform s-subset S,*

$$\mathbb{E}[\bar{Y}^4 \mid s] = 3\,\tau_4(s)\,V_2^2 + q(s)\,V_4,$$

where,

$$\tau_4(s) = p_2 - 2p_3 + p_4 = \frac{s(s-1)(d-s)(d-s-1)}{d(d-1)(d-2)(d-3)} \geq 0, \qquad q(s) = p_1 - 7p_2 + 12p_3 - 6p_4.$$

Moreover, writing $\tau(s) = p_1 - p_2 = \frac{s(d-s)}{d(d-1)}$ we get $\tau(s) - q(s) = 6\,\tau_4(s) \geq 0$, so $q(s) \leq \tau(s)$.

Proof. Let $\xi_i = \mathbf{1}[i \in S]$, so that $\bar{Y} = \sum_i \bar{v}_i \xi_i$ and, expanding the fourth power,

$$\bar{Y}^4 = \Big(\sum_i \bar{v}_i \xi_i\Big)^4 = \sum_{i,j,k,l} \bar{v}_i \bar{v}_j \bar{v}_k \bar{v}_l \, \xi_i \xi_j \xi_k \xi_l.$$

Taking expectations conditional on s,

$$\mathbb{E}[\bar{Y}^4 \mid s] = \sum_{i,j,k,l} \bar{v}_i \bar{v}_j \bar{v}_k \bar{v}_l \, \mathbb{E}[\xi_i \xi_j \xi_k \xi_l \mid s].$$

Since $\xi_i \in \{0,1\}$ satisfies $\xi_i^2 = \xi_i$, the value $\mathbb{E}[\xi_i \xi_j \xi_k \xi_l \mid s]$ depends only on the number m of *distinct* indices among i, j, k, l, and equals p_m. Grouping the ordered 4-tuples by their coincidence pattern (§A.8 carries out this grouping and the reductions below in full detail) gives

$$\mathbb{E}[\bar{Y}^4 \mid s] = p_1 V_4 + 4p_2 \sum_{i \neq j} \bar{v}_i^3 \bar{v}_j + 3p_2 \sum_{i \neq j} \bar{v}_i^2 \bar{v}_j^2 + 6p_3 \sum_{\substack{i,j,k \\ \text{distinct}}} \bar{v}_i^2 \bar{v}_j \bar{v}_k + p_4 \sum_{\substack{i,j,k,l \\ \text{distinct}}} \bar{v}_i \bar{v}_j \bar{v}_k \bar{v}_l.$$

Using $\sum_i \bar{v}_i = 0$, the power-sum reductions

$$\sum_{i \neq j} \bar{v}_i^3 \bar{v}_j = -V_4, \qquad \sum_{i \neq j} \bar{v}_i^2 \bar{v}_j^2 = V_2^2 - V_4,$$

$$\sum_{\substack{i,j,k \\ \text{distinct}}} \bar{v}_i^2 \bar{v}_j \bar{v}_k = 2V_4 - V_2^2, \qquad \sum_{\text{distinct}} \bar{v}_i \bar{v}_j \bar{v}_k \bar{v}_l = 3V_2^2 - 6V_4$$

(derived in §A.8) give, after collecting the coefficients of V_2^2 and V_4,

$$\mathbb{E}[\bar{Y}^4 \mid s] = 3(p_2 - 2p_3 + p_4) V_2^2 + (p_1 - 7p_2 + 12p_3 - 6p_4) V_4.$$

The closed form for $\tau_4 = p_2 - 2p_3 + p_4$ follows by clearing denominators: $p_2 - 2p_3 + p_4 = \frac{s(s-1)}{d(d-1)} \cdot B$ with $B = 1 - 2\frac{s-2}{d-2} + \frac{(s-2)(s-3)}{(d-2)(d-3)} = \frac{(d-s)(d-s-1)}{(d-2)(d-3)}$, the last equality by expanding the numerator to $(d-s)(d-s-1)$. Finally $\tau - q = (p_1 - p_2) - (p_1 - 7p_2 + 12p_3 - 6p_4) = 6(p_2 - 2p_3 + p_4) = 6\tau_4 \geq 0$. □

We simplify the terms in an upper bound. Since $V_4 \leq V_2^2$ (expand $(\sum_i \bar{v}_i^2)^2 = \sum_i \bar{v}_i^4 + \sum_{i \neq j} \bar{v}_i^2 \bar{v}_j^2 \geq \sum_i \bar{v}_i^4$) and, by Lemma 2, $q(s) \leq \tau(s)$, we have

$$\mathbb{E}[\bar{Y}^4 \mid s] = 3\tau_4(s) V_2^2 + q(s) V_4 \leq 3\tau_4(s) V_2^2 + \tau(s) V_4 \leq \big(3\tau_4(s) + \tau(s)\big) V_2^2.$$

Taking the expectation over s via the tower property $\mathbb{E}[\bar{Y}^4] = \mathbb{E}_s[\mathbb{E}[\bar{Y}^4 \mid s]]$,

$$\mathbb{E}[\bar{Y}^4] \leq \big(3\,\mathbb{E}_s[\tau_4] + \mathbb{E}_s[\tau]\big) V_2^2.$$

A.6 Expectation of the τ terms

Recall from Lemma 2 that $\tau(s) = \frac{s(d-s)}{d(d-1)}$ and $\tau_4(s) = \frac{s(s-1)(d-s)(d-s-1)}{d(d-1)(d-2)(d-3)}$. Under the Shapley kernel,

$$P(|x| = s) = \frac{1}{Z} \frac{1}{s(d-s)}, \qquad Z = \sum_{s=1}^{d-1} \frac{1}{s(d-s)} = \frac{2H_{d-1}}{d}, \tag{12}$$

Lemma 3 (Shapley averages of τ and τ_4)**.** *Under equation 12,*

$$\mathbb{E}_s[\tau] = \frac{1}{2H_{d-1}}, \qquad \mathbb{E}_s[\tau_4] = \frac{1}{12H_{d-1}}.$$

Proof. By equation 12, $\mathbb{E}_s[\tau] = \frac{1}{Z} \sum_{s=1}^{d-1} \frac{1}{s(d-s)} \frac{s(d-s)}{d(d-1)} = \frac{d-1}{Z\, d(d-1)} = \frac{1}{Zd} = \frac{1}{2H_{d-1}}$. For τ_4, the factor $\frac{1}{s(d-s)}$ cancels the leading $s(d-s)$, leaving

$$\mathbb{E}_s[\tau_4] = \frac{1}{Z\, d(d-1)(d-2)(d-3)} \sum_{s=1}^{d-1} (s-1)(d-s-1) = \frac{(d-1)(d-2)(d-3)/6}{Z\, d(d-1)(d-2)(d-3)} = \frac{1}{6Zd} = \frac{1}{12H_{d-1}},$$

using $\sum_{j=0}^{d-2} j(d-2-j) = \frac{(d-1)(d-2)(d-3)}{6}$. □

A.7 PUTTING IT ALL TOGETHER

We now put all the pieces together. From equation 11 we get:

$$\sup_{v^\top \Sigma v=1} \mathbb{E}[(z_c^\top v)^4] \leq 128\,\mathbb{E}[\bar{Y}^4] + O(1). \tag{13}$$

Then from Section A.5 we get $\mathbb{E}[\bar{Y}^4] \leq \big(3\,\mathbb{E}_s[\tau_4] + \mathbb{E}_s[\tau]\big)V_2^2$ and from Section A.3 we have $V_2^2 \leq \frac{1}{\alpha^2} = \frac{1}{4}H_{d-1}^2$. Finally from Section A.5 we get the following:

$$\big(3\,\mathbb{E}_s[\tau_4] + \mathbb{E}_s[\tau]\big)V_2^2 = \Big(\frac{3}{12H_{d-1}} + \frac{1}{2H_{d-1}}\Big)V_2^2 = \frac{3}{4H_{d-1}}\,V_2^2.$$

Thus:

$$\sup_{v^\top \Sigma v=1} \mathbb{E}[(z_c^\top v)^4] \leq O(1) + 128(\frac{3}{4H_{d-1}}) \cdot (\tfrac{1}{4}H_{d-1}^2) = O(1) + 24H_{d-1} = O(1) + O(\log d)$$

Therefore we get $C \leq O(\log d)$.

A.8 DERIVATION OF THE EXACT FOURTH MOMENT

This subsection expands the grouping and reduction steps compressed in the proof of Lemma 2. Throughout, S is a uniform s-subset of $\{1, \ldots, d\}$, $\xi_i = \mathbf{1}[i \in S]$, and the population $\{\bar{v}_i\}$ is centred ($\sum_i \bar{v}_i = 0$).

(i) The expectation depends only on the number of distinct indices. Since $\xi_i \in \{0, 1\}$ we have $\xi_i^2 = \xi_i$, so any repeated index in a product $\xi_i\xi_j\xi_k\xi_l$ collapses. Consequently, if the tuple (i, j, k, l) takes exactly m distinct values, then $\mathbb{E}[\xi_i\xi_j\xi_k\xi_l \mid s] = p_m$, regardless of the multiplicities.

(ii) Grouping the quadruple sum by equality pattern. For a tuple (i, j, k, l), record which of the four positions carry equal index values – this is a set partition π of the four positions $\{1, 2, 3, 4\}$, with two positions in the same block iff they carry the same index. Summing over all tuples with a *fixed* pattern π having blocks $B_1, \ldots, B_m$ means assigning a value to each block, with values in distinct blocks forced to be *distinct*. By (i), every such tuple has expectation p_m, so the contribution of π is

$$p_m \sum_{\substack{a_1,\ldots,a_m \\ \text{all distinct}}} \prod_{t=1}^{m} \bar{v}_{a_t}^{|B_t|}.$$

Partitions of the same *type* (same multiset of block sizes) contribute equally, so we group by type and multiply by the number of partitions of that type.

The five patterns. The set partitions of $\{1, 2, 3, 4\}$ fall into five types, totalling the Bell number $B_4 = 15$:

type (block sizes)	# partitions	m	moment	block-value sum
$[4]$	1	1	p_1	$\sum_a \bar{v}_a^4 = V_4$
$[3, 1]$	4	2	p_2	$\sum_{a \neq b} \bar{v}_a^3 \bar{v}_b$
$[2, 2]$	3	2	p_2	$\sum_{a \neq b} \bar{v}_a^2 \bar{v}_b^2$
$[2, 1, 1]$	6	3	p_3	$\sum_{a,b,c \text{ dist.}} \bar{v}_a^2 \bar{v}_b \bar{v}_c$
$[1, 1, 1, 1]$	1	4	p_4	$\sum_{a,b,c,d \text{ dist.}} \bar{v}_a \bar{v}_b \bar{v}_c \bar{v}_d$

(Counts: $[4]$ one way; $[3,1]$ choose the singleton position, $\binom{4}{1}=4$; $[2,2]$ pair up four positions, 3 ways; $[2,1,1]$ choose the doubled pair of positions, $\binom{4}{2}=6$; $[1,1,1,1]$ one way.) Hence

$$\mathbb{E}[\bar{Y}^4 \mid s] = p_1 V_4 + 4p_2 \sum_{a\neq b} \bar{v}_a^3 \bar{v}_b + 3p_2 \sum_{a\neq b} \bar{v}_a^2 \bar{v}_b^2 + 6p_3 \sum_{\text{dist.}} \bar{v}_a^2 \bar{v}_b \bar{v}_c + p_4 \sum_{\text{dist.}} \bar{v}_a \bar{v}_b \bar{v}_c \bar{v}_d. \tag{14}$$

Reducing each block-value sum via $\sum_i \bar{v}_i = 0$. Each restricted sum is converted to V_2, V_4 by inclusion–exclusion over index coincidences.

$[3,1]$*:*

$$\sum_{a\neq b} \bar{v}_a^3 \bar{v}_b = \Big(\sum_a \bar{v}_a^3\Big)\Big(\sum_b \bar{v}_b\Big) - \sum_a \bar{v}_a^4 = 0 - V_4 = -V_4.$$

$[2,2]$*:*

$$\sum_{a\neq b} \bar{v}_a^2 \bar{v}_b^2 = \Big(\sum_a \bar{v}_a^2\Big)\Big(\sum_b \bar{v}_b^2\Big) - \sum_a \bar{v}_a^4 = V_2^2 - V_4.$$

$[2,1,1]$*:* Start from the unrestricted sum $U = \sum_{a,b,c} \bar{v}_a^2 \bar{v}_b \bar{v}_c = V_2\big(\sum_c \bar{v}_c\big)\big(\sum_b \bar{v}_b\big)/\ldots$ More directly, split $U = V_2 \cdot 0 \cdot 0 = 0$ by the equality pattern of (a,b,c) (where a carries the square):

$$0 = U = \underbrace{\sum_{\text{dist.}} \bar{v}_a^2 \bar{v}_b \bar{v}_c}_{=:T} + \underbrace{\sum_{a\neq c} \bar{v}_a^3 \bar{v}_c}_{a=b} + \underbrace{\sum_{a\neq b} \bar{v}_a^3 \bar{v}_b}_{a=c} + \underbrace{\sum_{a\neq b} \bar{v}_a^2 \bar{v}_b^2}_{b=c} + \underbrace{\sum_a \bar{v}_a^4}_{a=b=c}.$$

Substituting the values already found ($-V_4$, $-V_4$, $V_2^2 - V_4$, V_4),

$$0 = T + (-V_4) + (-V_4) + (V_2^2 - V_4) + V_4 = T + V_2^2 - 2V_4 \implies T = 2V_4 - V_2^2.$$

$[1,1,1,1]$*:* The ordered all-distinct sum is $4!$ times the elementary symmetric polynomial e_4. Newton's identity gives

$$\sum_{a,b,c,d \text{ dist.}} \bar{v}_a \bar{v}_b \bar{v}_c \bar{v}_d = 24\, e_4 = V_1^4 - 6V_1^2 V_2 + 3V_2^2 + 8V_1 V_3 - 6V_4,$$

where $V_1 = \sum_i \bar{v}_i = 0$; this collapses to $3V_2^2 - 6V_4$.

Collecting. Insert the four reductions into equation 14:

$$\mathbb{E}[\bar{Y}^4 \mid s] = p_1 V_4 + 4p_2(-V_4) + 3p_2(V_2^2 - V_4) + 6p_3(2V_4 - V_2^2) + p_4(3V_2^2 - 6V_4).$$

Collecting the coefficient of V_2^2, $3p_2 - 6p_3 + 3p_4 = 3(p_2 - 2p_3 + p_4)$, and of V_4, $p_1 - 4p_2 - 3p_2 + 12p_3 - 6p_4 = p_1 - 7p_2 + 12p_3 - 6p_4$, gives exactly

$$\mathbb{E}[\bar{Y}^4 \mid s] = 3(p_2 - 2p_3 + p_4)V_2^2 + (p_1 - 7p_2 + 12p_3 - 6p_4)V_4,$$

matching Lemma 2.

A.9 From Explained Variance to $R^2_{\max}$

Let

$$V_y = \mathrm{Var}(y) > 0, \qquad \mu_4 = \mathbb{E}[(y - \mathbb{E}[y])^4],$$

and let

$$\widehat{V}_y = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

be the unbiased sample variance. Its variance is

$$\mathrm{Var}(\widehat{V}_y) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1} V_y^2\right) \leq \frac{\mu_4}{n}.$$

Because

$$r = \frac{y - \mathbb{E}[y]}{\sqrt{V_y}}$$

and $\mathbb{E}[r^4] \leq \kappa$, we have

$$\mu_4 \leq \kappa V_y^2.$$

Therefore,

$$\mathrm{Var}(\widehat{V}_y) \leq \frac{\kappa V_y^2}{n}.$$

By Chebyshev's inequality, for every $\eta > 0$,

$$\Pr\left(\left|\frac{\widehat{V}_y}{V_y} - 1\right| > \eta\right) \leq \frac{\kappa}{n\eta^2}.$$

Setting

$$\eta = \sqrt{\frac{2\kappa}{n\tau}}$$

shows that, with probability at least $1 - \tau/2$,

$$\left|\frac{\widehat{V}_y}{V_y} - 1\right| \leq \sqrt{\frac{2\kappa}{n\tau}}.$$

B Experiment Design Details

This section provides additional details about the design of the experiments.

B.1 SHAP Details

To sample data points from the local dataset we use the Captum Kokhlikyan et al. (2020) library's KernelSHAP implementation as it samples the local dataset according to the shapley distribution instead of weighting the regression problem according to the shapley weights. The perturbation mapping h_x is done by replacing features that are turned off with 0 which is equivalent to mean replacement as the data is standardized. The outputs y in the local dataset are the raw output of the model centered as is done in the original KernelSHAP Lundberg & Lee (2017a).

B.2 Real-World Dataset Details

In this section we describe the details of the datasets used for the experiments and details about how we compute the ground truth values for those datasets. The real world datasets are shown in Table 2.

Name	d	Description	Link	Citation
Australian	14	Predict whether to accept a credit card application.	Link	Quinlan (1987)
Spectfheart	44	Predict whether a cardiac image is normal or abnormal.	Link	Kurgan et al. (2001)
Sonar	60	Distinguish rocks from mines in sonar data.	Link	Gorman & Sejnowski (1988)
Tuandromd	240	Distinguish malware from benign programs.	Link	Borah et al. (2020)
Parkinsons	753	Predict whether the patient has parkinsons.	Link	Sakar et al. (2019)
Qsar Toxic	1024	Predict whether a compound is toxic.	Link	Ballabio et al. (2019)

Table 2: Real-World Datasets

To estimate the ground truth additive interpretability on the real world datasets we compute a KernelSHAP estimate of the optimal explanation using n samples and test its performance using t samples. Below are the exact numbers used shown as n+t. For German and Sonar 50K+10K samples are used to estimate the ground truth, for Tuandromd 100K+20K and for Parkinsons and Qsar Toxic 200K+25K samples are used. These sample sizes are determined by increasing the number of samples until the estimate converges determined by the criteria $|v_t - v_{t-3}|/v_t < 0.05$ where v_t is the value at

Dataset	Excess Kurtosis
Australian	1.45 ± 2.57
Spectfheart	1.76 ± 2.83
Sonar	1.43 ± 0.87
Tuandromd	2.38 ± 1.00
Parkinsons	0.14 ± 1.29
QSAR Toxic	-0.05 ± 0.77
Total	1.18 ± 1.96

Table 3: Excess Kurtosis of the Neural Network on the local dataset of data point from the real datasets (mean $\pm$ std).

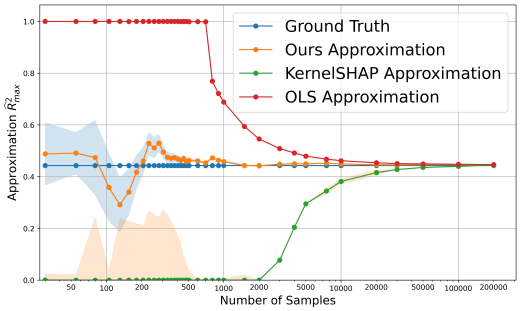


Figure 3: Convergence of the ground truth estimate on an example datapoint from the Parkinsons dataset. The x-axis shows the number of samples, while the y-axis shows the estimated value. The figure shows how KernelSHAP converges to the value used as the ground truth value, with our estimator and OLS as reference.

the t-th sample step and $n > d$ since the estimate does not improve for $n < d$. See Figure 3 for an example.

The instances of the datasets used for the experiments are chosen at random and the reruns are seeded with the index of the given run. The excess kurtosis of the model output on the datasets is shown in Table 3.

B.3 Synthetic Function Details

For the synthetic datasets we randomly generate multilinear polynomials and a data point to explain them on. Multilinear polynomials have an exact formula for the shapley values allowing us to compute the exact ground truth. Since, the exact formula is exponential for general multilinear polynomials we reduce the computation to linear time by keeping them sparse. Specifically we only allow interaction terms up to the fifth order and at each order only allow $0.8 * d$ interaction terms.

B.4 Model Details

Dataset	Neural Network Accuracy
Australian	0.88
Spectfheart	0.70
Sonar	0.72
Tuandromd	0.73
Parkinsons	0.67
Qsar Toxicity	0.77

Table 4: Neural Network performance (accuracy) on each real-world dataset.

The model being explained for real datasets is a neural network. The neural network consists of 3 hidden layers with width equal to half the input dimension. The neural network is trained with the ADAM optimizer Kingma & Ba (2015) and cross-entropy loss for 100 epochs with early stopping when the validation accruacy is above 0.8. We standarize the data and use a $80/20$ split of balanced data to train and evaluate the model. Table 4 shows the accuracy of the trained model on the given datasets.

B.5 Hyperparameter settings for baseline estimators

In this section, we describe the hyperparameters used for the baseline methods. For all methods we standardize the local input as described in Section A.1.

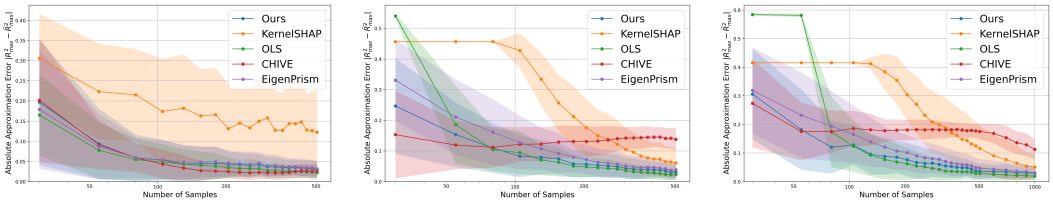
For KernelSHAP we use a $80/20$ split of the available samples for training and testing. Since the least squares solution for $\widehat{\theta}$ used in KernelSHAP is not defined when $n < d$ we use regularised regression in this case.

For Ordinary Least Squares the estimate is similarly not defined when $n < d$. We again use regularised regression for $\widehat{\theta}$ and replace the denominator with n in the estimator, following the biased estimator described by Kong & Valiant (2018).

For CHIVE the additional hyperparameters are the initial λ which is set to 2.01 and the number of iterations which are set to 100 used for the scaled lasso estimation.

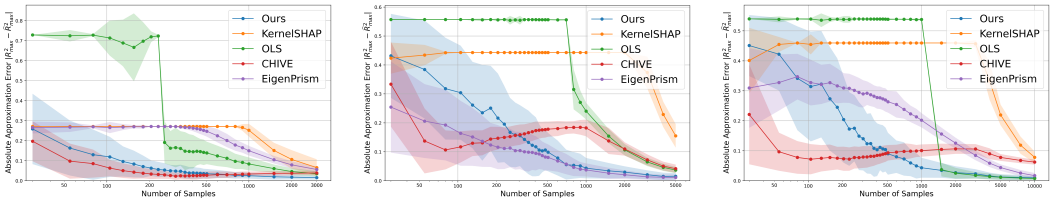
C Full Experiment Results

In this section we show additional experiments we could not fit into main body's experiment section. Figure 4 and Figure 5 show example convergence curves for low and high dimensional datasets respectively.



(a) Australian (d = 14) (b) Spectfheart (d = 44) (c) Sonar (d = 60)

Figure 4: Convergence comparison across low-dimensional datasets (shown in log scale). The x-axis shows increasing number of samples, while the y-axis shows the absolute error of the estimator. The line shows the mean error across 50 runs, while the shaded area shows the standard deviation.



(a) Tuandromd (d = 240) (b) Parkinsons (d = 753) (c) Qsar Toxic (d = 1024)

Figure 5: Convergence comparison across high-dimensional datasets (shown in log scale). The x-axis shows increasing number of samples, while the y-axis shows the absolute error of the estimator. The line shows the mean error across 50 runs, while the shaded area shows the standard deviation.

Tables 5 and 6 show the sample complexity on the real datasets for the thresholds 0.1 and 0.05 respectively. Tables 7 and 8 show the sample complexity on the synthetic datasets for the thresholds 0.1 and 0.05 respectively.

	Australian ($d = 14$)	Spectfheart ($d = 44$)	Sonar ($d = 60$)	Tuandromd ($d = 240$)	Parkinsons ($d = 753$)	Qsart ($d = 1024$)
Ours	157 ± 80	169 ± 84	202 ± 75	$\mathbf{227 \pm 124}$	$\mathbf{661 \pm 283}$	$\mathbf{383 \pm 139}$
KernelSHAP	203 ± 122	508 ± 318	498 ± 144	1835 ± 112	7156 ± 771	8281 ± 688
OLS	$\mathbf{71 \pm 24}$	$\mathbf{124 \pm 59}$	$\mathbf{121 \pm 25}$	825 ± 240	1952 ± 557	1163 ± 81
CHIVE	149 ± 89	680 ± 472	1351 ± 209	301 ± 471	2101 ± 791	1025 ± 1878
EigenPrism	166 ± 69	238 ± 53	201 ± 67	1511 ± 210	1502 ± 1192	4023 ± 792

Table 5: Number of samples needed to reach an absolute-error threshold of 0.1, averaged across 10 data points per dataset (mean $\pm$ std).

	Australian ($d = 14$)	Spectfheart ($d = 44$)	Sonar ($d = 60$)	Tuandromd ($d = 240$)	Parkinsons ($d = 753$)	Qsart ($d = 1024$)
Ours	272 ± 130	357 ± 283	431 ± 247	$\mathbf{383 \pm 249}$	$\mathbf{1040 \pm 838}$	$\mathbf{551 \pm 164}$
KernelSHAP	294 ± 184	864 ± 596	930 ± 233	3142 ± 724	9708 ± 604	$> s_{max}$
OLS	$\mathbf{97 \pm 40}$	$\mathbf{201 \pm 86}$	$\mathbf{181 \pm 82}$	1685 ± 543	3557 ± 1691	1832 ± 1285
CHIVE	258 ± 213	1148 ± 849	2766 ± 809	1112 ± 2025	5302 ± 2830	2082 ± 3801
EigenPrism	323 ± 148	386 ± 158	439 ± 265	2917 ± 728	2928 ± 2576	7821 ± 949

Table 6: Number of samples needed to reach an absolute-error threshold of 0.05, averaged across 10 data points per dataset (mean $\pm$ std).

	Synthetic ($d = 50$)	Synthetic ($d = 100$)	Synthetic ($d = 250$)	Synthetic ($d = 500$)	Synthetic ($d = 1000$)	Synthetic ($d = 2000$)
Ours	212 ± 72	301 ± 184	$\mathbf{366 \pm 62}$	$\mathbf{609 \pm 669}$	$\mathbf{581 \pm 438}$	$\mathbf{1241 \pm 885}$
KernelSHAP	254 ± 104	479 ± 108	1570 ± 579	3304 ± 1758	4749 ± 1252	9499 ± 852
OLS	$\mathbf{101 \pm 48}$	$\mathbf{159 \pm 32}$	579 ± 271	1313 ± 1047	1067 ± 1300	5642 ± 3954
CHIVE	503 ± 403	1060 ± 590	4655 ± 673	4736 ± 4489	4019 ± 4883	8005 ± 3988
EigenPrism	229 ± 91	397 ± 171	753 ± 421	2628 ± 1036	5749 ± 1901	7255 ± 3683

Table 7: Number of samples needed to reach an absolute-error threshold of 0.1, averaged across 10 models per dimension (mean $\pm$ std).

	Synthetic ($d = 50$)	Synthetic ($d = 100$)	Synthetic ($d = 250$)	Synthetic ($d = 500$)	Synthetic ($d = 1000$)	Synthetic ($d = 2000$)
Ours	239 ± 74	311 ± 177	$\mathbf{427 \pm 114}$	$\mathbf{663 \pm 657}$	$\mathbf{640 \pm 402}$	$\mathbf{1323 \pm 851}$
KernelSHAP	436 ± 223	808 ± 197	2874 ± 1200	5317 ± 2739	7446 ± 1984	$> s_{max}$
OLS	$\mathbf{125 \pm 50}$	$\mathbf{230 \pm 50}$	1105 ± 462	2270 ± 1419	3314 ± 1191	8372 ± 2108
CHIVE	708 ± 662	1611 ± 898	8421 ± 1145	6023 ± 4870	4030 ± 4874	8015 ± 3968
EigenPrism	299 ± 82	554 ± 253	1158 ± 618	4234 ± 1875	8520 ± 2170	8070 ± 3859

Table 8: Number of samples needed to reach an absolute-error threshold of 0.05, averaged across 10 models per dimension (mean $\pm$ std).