BESTSHAP: BEST-OF-MANY ADAPTIVE APPROXIMATIONS FOR THE SHAPLEY VALUE

Anonymous authors
Paper under double-blind review

ABSTRACT

The Shapley value has become a popular concept in attribution problems because it is uniquely characterized by a specified set of axioms. However, computing the Shapley value for a general characteristic function requires exponentially many queries, motivating extensive research into randomized approximation algorithms. Although several randomized algorithms achieve the best-known query complexity for a specified approximation guarantee, none consistently outperforms the others in practice. In other words, without access to the ground-truth Shapley values, it remains unclear which estimate is the most accurate. In this work, we show that identifying the potentially best estimate can come for free, without requiring access to the ground-truth Shapley values. As a first step, we introduce a class of provably adaptive randomized algorithms that retain the best-known query complexity while requiring only linear memory. These algorithms differ in their choice of control variates, and the effectiveness of the adaptivity associated with each variate depends on the characteristic function used. In particular, this class naturally generalizes the linear-memory KernelSHAP framework. The estimates produced by these algorithms can be obtained simultaneously using the same sequence of sampled subsets. Moreover, the mean squared errors of these estimates, which measure their expected approximation quality, can be approximated in parallel without increasing the time or memory complexity. These estimates of the approximation error thus enable us to select the potentially best candidate. Theoretically, the accuracy of this selection improves as the number of sampled subsets increases. Our theoretical results are supported by empirical evaluations. Furthermore, our proposed algorithm, BestSHAP, consistently outperforms existing linear-memory baselines.

1 INTRODUCTION

Originating in economics, cooperative game theory studies how to justifiably allocate the value generated by a cooperative game among n players (Grabisch, 2016). A cooperative game is represented by a characteristic function $U_n\colon 2^{[n]} \to \mathbb{R}$, where $[n] \coloneqq \{1, 2, \ldots, n\}$, and $U(S)$ denotes the value generated by coalition $S \subseteq [n]$. For convenience, the subscript n of U_n is used to indicate the number of participants, and we denote the set all possible characteristic functions by $\mathcal{G} \coloneqq \bigcup_{n\geq 2}\{U_n\colon 2^{[n]} \to \mathbb{R}\}$. Specifically, an attribution method is a mapping ϕ that maps each $U_n \in \mathcal{G}$ into a contribution vector $\boldsymbol{\phi} \coloneqq \phi(U_n) \in \mathbb{R}^n$, where ϕ_i quantifies the contribution of the i-th player. In this field, the Shapley value (Shapley, 1953) is the most well-known attribution method. Its popularity is mainly due to that it uniquely satisfies the following axioms:

Linearity: $\phi(c \cdot U_n + U'_n) = c \cdot \phi(U_n) + \phi(U'_n)$ for every $U_n, U'_n \in \mathcal{G}$ and $c \in \mathbb{R}$;

Consistency: Suppose there exists $c \in \mathbb{R}$ such that $U_{n+1}(S) = U_n(S \cap [n]) + c \cdot [\![n+1 \in S]\!]$ for every $S \subseteq [n+1]$. Then, (i) $\phi_{n+1}(U_{n+1}) = c$, and (ii) $\phi_i(U_n) = \phi_i(U_{n+1})$ for every $i \in [n]$. Here, $[\![\cdot]\!]$ denotes the indicator function;

Equal Treatment: For some $i, j \in [n]$, if $U_n(S \cup \{i\}) = U_n(S \cup \{j\})$ for every $S \subseteq [n] \setminus \{i, j\}$, then $\phi_i(U_n) = \phi_j(U_n)$;

Efficiency: $\sum_{i\in[n]} \phi_i(U_n) = U_n([n]) - U_n(\emptyset)$ for every $U_n \in \mathcal{G}$.

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 recent years, the Shapley value has been applied to various AI attribution problems, with prominent examples including data attribution and feature attribution. In the former, (e.g., He et al., 2024; Wang et al., 2025; Tian et al., 2026; Lu et al., 2026), individual training data points or groups of them constitute the set of players, and $U(S)$ is defined as the performance of a model trained or fine-tuned on the subset S. In the latter (e.g., Lundberg & Lee, 2017; Lundberg et al., 2020; Frye et al., 2021; Butler et al., 2025), features are viewed as players, and $U(S)$ is defined as the output of a trained model for a particular instance when the features outside S are treated as missing.

Specifically, while using the Shapley value, the contribution of the i-th player is defined as

$$\phi_i(U_n) \coloneqq \sum_{S \subseteq [n] \setminus \{i\}} p_{|S|+1} \cdot [U_n(S \cup \{i\}) - U_n(U)] \text{ where } p_s \coloneqq \frac{1}{n} \binom{n-1}{s-1}^{-1}. \tag{1}$$

Clearly, computing the contribution vector $\boldsymbol{\phi} \coloneqq \phi(U_n)$ requires exponentially many evaluations of U_n, making exact computation intractable in general. This motivates a recent line of research that explores how to efficiently approximate the Shapley value without imposing restrictive structural assumptions on the characteristic function U (e.g., Jia et al., 2019; Covert & Lee, 2021; Zhang et al., 2023; Kolpaczki et al., 2024; Fumagalli et al., 2023; Li & Yu, 2024a;b; Musco & Witter, 2025; Chen et al., 2025; Li et al., 2026a). Specifically, theoretical guarantee is based on the asymptotic query complexities required to achieve $P(\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2 \geq \epsilon) \leq \delta$, while assuming that $\max_{S \subseteq [n]} |U_n(S)| \leq C$ as $n \to \infty$. Very recently, Li et al. (2026a) established that, by selecting an appropriate sampling distribution within the linear-memory KernelSHAP framework, several resulting variants achieve the best-known query complexity $O\left(\frac{n}{\epsilon^2} \log \frac{1}{\delta}\right)$, which is linear in the number of players n. In particular, their new method Adalina was derived by shifting U to minimize the mean squared error (MSE):

$$\text{MSE} \coloneqq \mathbb{E}[\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2^2],$$

leading to the first provably adaptive variant to automatically reduce the MSE for each characteristic function U using a constant control variate. However, our experimental results indicate that none of these variants consistently outperforms the others. It raises the following question:

Without access to the ground-truth Shapley values, how can we determine which variant is the best choice for a given characteristic function U?

In this work, we show that such selection is possible without incurring any additional cost in terms of memory or query complexity.

Our theoretical contributions. We first introduce a class of *provably* adaptive randomized algorithms that includes Adalina (Li et al., 2026a). These algorithms differ in their choice of control variates for minimizing the MSE; for Adalina, the control variate corresponds to the constant characteristic function $U \equiv 1$. In particular, we theoretically establish a sufficient condition on the control variate under which the resulting randomized algorithm retains the best-known query complexity. This condition provides a foundation for designing many new adaptive randomized algorithms with linear memory and query complexities.

Intuitively, no single adaptive randomized algorithm consistently achieves the minimum MSE for all characteristic functions. Since the MSE measures the expected squared approximation error w.r.t. the ground-truth Shapley values, we use it as the decisive criterion for selecting the best estimate among those produced by the different adaptive randomized algorithms. Importantly, we can approximate a surrogate for the MSE without access to the ground-truth $\boldsymbol{\phi}$. Moreover, this additional approximation does not increase the memory or query complexity. We theoretically establish that, when the number of sampled subsets is sufficiently large, the algorithm with the smallest estimated surrogate value is highly likely to be the one with the smallest MSE.

Our empirical contributions. We consolidate our theoretical results in the proposed BestSHAP (best of many) algorithm. BestSHAP generates a sequence of subsets *once* to obtain estimates from many adaptive randomized algorithms, together with estimates of their corresponding MSE surrogates. It then selects the estimate with the smallest estimated surrogate value. Our experimental results verify that these estimated surrogate values align well with the empirical approximation errors. Moreover, BestSHAP consistently outperforms all the linear-memory baselines. We will release our code once this submission is de-anonymized.

2 Related Works

Least-squares-based randomized algorithms. A line of work has developed randomized algorithms based on solving least-squares problems (Lundberg & Lee, 2017). Specifically, as first observed by Charnes et al. (1988, Theorem 3), the Shapley value is exactly the coefficient vector $\mathbf{x}$ of the uniquely optimal V to

$$\underset{V \in \mathcal{L} \cap \mathcal{E}_U}{\text{minimize}} \sum_{\emptyset \subsetneq S \subsetneq [n]} w_{|S|} \cdot (U(S) - U(\emptyset) - V(S))$$

$$\text{where } \mathcal{L} \coloneqq \{V \mid V(S) = \sum_{i \in S} x_i\} \text{ and } \mathcal{E}_U \coloneqq \{V \mid V([n]) = U([n]) \wedge V(\emptyset) = 0\}.$$

Here, $w_s = \binom{n-2}{s-1}^{-1}$. In other words, the Shapley value of U is the coefficient vector $\mathbf{x}$ associated with the orthogonal projection of U onto the affine space $\mathcal{L} \cap \mathcal{E}_U$, i.e., $\phi(U) = \text{Proj}_{\mathcal{L} \cap \mathcal{E}_U}(U)$.[1] The most recent related works (Fumagalli et al., 2026b;a) essentially extend $\mathcal{L}$ to a larger linear space $\overline{\mathcal{L}}$, and then employ the basic fact that

$$\phi(V^*) = \text{Proj}_{\mathcal{L} \cap \mathcal{E}_U}\left(\text{Proj}_{\overline{\mathcal{L}} \cap \mathcal{E}_U}(U)\right) = \text{Proj}_{\mathcal{L} \cap \mathcal{E}_U}(U) = \phi(U),$$

where $V^* \coloneqq \text{Proj}_{\overline{\mathcal{L}} \cap \mathcal{E}_U}(U)$ is the best approximation to U in $\overline{\mathcal{L}} \cap \mathcal{E}_U$. These methods typically require $\Theta(d^2)$ memory, where $d \geq n$, primarily to store the $d \times d$ matrix required to solve the corresponding least-squares problem. They also incur an additional $O(d^3)$ computational cost to invert this matrix when solving the stored approximate least-squares problem, which can be prohibitive for downstream tasks where n is large (He et al., 2024; Tian et al., 2026). In contrast, linear-memory randomized algorithms require only $O(n)$ overhead to generate an estimate from stored information. To our knowledge, the best-known query complexity for these least-squares-based algorithms is $O\left(\frac{n}{\epsilon^2} \log n + \frac{n}{\epsilon^2 \delta}\right)$ (Chen et al., 2025, Theorem 2.2),[2] which is worse than the best-known query complexity $\left(\frac{n}{\epsilon^2} \log \frac{1}{\delta}\right)$ for linear-memory algorithms. We note that different randomized algorithms may be better suited to characteristic functions with different structures, and that these query-complexity results provide only upper bounds.[3]

Fitting a surrogate characteristic function W to U. A complementary technique for *independently* improving approximation quality is to fit a surrogate characteristic function W to U (Witter et al., 2025; Liu et al., 2026). Specifically, W is selected so that $\boldsymbol{\phi}(W)$ can be computed in polynomial time. Examples include $W \in \mathcal{L}$, W incorporating higher-order interaction terms (Fumagalli et al., 2026a, Eqs. (3) and (14)), and W based on interventional gradient-boosting decision trees (Lundberg et al., 2020, Algorithm 3). Since $\phi(U) = \phi(W) + \phi(U - W)$, the sampled subsets can then be reused to approximate the residual characteristic function $U - W$, yielding the estimator,

$$\widehat{\boldsymbol{\phi}}(U) \coloneqq \boldsymbol{\phi}(W) + \widehat{\boldsymbol{\phi}}(U - W).$$

To our knowledge, existing related query-complexity analyses account only for the residual, as the extent to which W fits U was circumvented. Intuitively, the difficulty may lie in the fact that a limited number of sampled subsets cannot characterize an arbitrary U, unless the unseen subsets are highly correlated with the observed ones. On a high level, the randomized algorithms introduced in this work can be viewed as further improving the approximation of the residual.

[1] This least-squares connection continues to hold if w_0 and w_n are assigned arbitrary positive weights, yielding a valid inner-product space.

[2] To our knowledge, the technique used to remove the $\log n$ factor for linear-memory algorithms is not applicable to least-squares-based algorithms.

[3] Since $\sum_{S \subseteq [n] \setminus \{i\}} p_{|S|+1} = 1$, Eq. (1) can be rewritten as $\phi_i(U) = \mathbb{E}_{S \subseteq [n] \setminus \{i\}}[U(S \cup \{i\}) - U(S)]$. When $\text{Var}_{S \subseteq [n] \setminus \{i\}}(U(S \cup \{i\}) - U(S))$ is sufficiently small for every $i \in [n]$ (Jia et al., 2019), then averaging many independently sampled quantities of the form $U(S \cup \{i\}) - U(S)$ could outperform all existing state-of-the-art randomized algorithms, despite having a query complexity of $O(\frac{n^2}{\epsilon^2} \log \frac{n}{\delta})$ (Li & Yu, 2024a, Proposition 5).

3 BACKGROUND

Notation. For each subset S, the lowercase s is used to denote its cardinality $|S|$. In addition, bold letters indicate vectors or random variables. For example, $\mathbf{S}$ may represent a random subset. Each characteristic function $U\colon 2^{[n]} \to \mathbb{R}$ may be treated as a vector $\mathbf{u} \in \mathbb{R}^{2^n}$. In addition, $[\![\cdot]\!]$ denotes the indicator function.

We follow the linear-memory framework recently introduced by Li et al. (2026a), which bridges several previous works: SHAP-IQ (Fumagalli et al., 2023), OFA (Li & Yu, 2024b), MSR (Witter et al., 2025), and KernelSHAP (Chen et al., 2025). This framework begins by observing that the Shapley value, defined in Eq. (1), can be expressed as a sum of vectors,

$$\boldsymbol{\phi} = \boldsymbol{\varphi} + \mathbf{c} \;\text{ where }\; \mathbf{c} \coloneqq \frac{u_{[n]} - u_\emptyset}{n} \cdot \mathbf{1}_n, \tag{2}$$

$\boldsymbol{\varphi} \coloneqq \sum_{\emptyset \subsetneq S \subsetneq [n]} u_S \mathbf{w}_S$, $u_S \coloneqq U(S)$ and $(\mathbf{w}_S)_i = p_s [\![i \in S]\!] - p_{s+1} [\![i \notin S]\!]$. Accordingly, it is sufficient to focus on approximating $\boldsymbol{\varphi}$. Let $\mathbf{q} \in \mathbb{R}^{n-1}_{>0}$ be a sampling distribution over subset sizes $s \in [n-1]$, where q_s denotes the probability of sampling a subset of size s. Sampling according to $\mathbf{q}$ generates an independent sequence of subsets $\{S_t\}_{t=1}^T$. For a subset S of size s, define

$$\mathbf{z}_S = \frac{\binom{n}{s}}{q_s} \mathbf{w}_S \;\text{ so that }\; (\mathbf{z}_S)_i = \frac{1}{q_s} \left(\frac{[\![i \in S]\!]}{s} - \frac{[\![i \notin S]\!]}{n-s} \right), \; \forall\, i \in [n]. \tag{3}$$

Then, an unbiased estimate of $\boldsymbol{\phi}$ is

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\varphi}} + \mathbf{c} \;\text{ where }\; \widehat{\boldsymbol{\varphi}} \coloneqq \frac{1}{T} \sum_{t=1}^{T} u_{S_t} \cdot \mathbf{z}_{S_t}. \tag{4}$$

In particular, the query complexity of $\widehat{\boldsymbol{\phi}}$ is essentially $O\left(\frac{D(\mathbf{q}) \cdot n}{\epsilon} \log \frac{1}{\delta}\right)$, with the constant $D(\mathbf{q})$ governed by the sampling distribution $\mathbf{q}$. Minimizing $D(\mathbf{q})$ therefore yields the optimal choice

$$q_s^* \propto \frac{1}{\sqrt{s(n-s)}} \tag{5}$$

and $D \coloneqq D(\mathbf{q}^*) \in \Theta(1)$. A particularly remarkable property of $\mathbf{q}^*$ is that

$$\|\mathbf{z}_S\|_2^2 = nD, \quad \forall\, \emptyset \subsetneq S \subsetneq [n]. \tag{6}$$

Crucially, the squared 2-norm of the random *vector* $u_{\mathbf{S}} \cdot \mathbf{z}_{\mathbf{S}}$, which appears in Eq. (4), reduces to the random *scalar* $nD \cdot u_{\mathbf{S}}^2$. This property is key to developing *provably* adaptive randomized algorithms that use control variates to further reduce the MSE of $\widehat{\boldsymbol{\phi}}$. Due to its attractive theoretical properties, $\mathbf{q}^*$ is assumed throughout the paper unless stated otherwise.

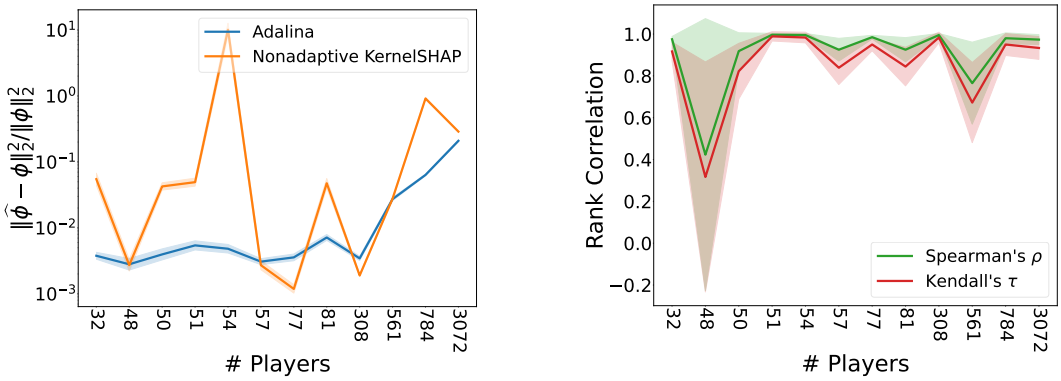
The use of $\mathbf{q}^*$ leads to two linear-memory KernelSHAP variants with linear query complexity: Adalina, which is adaptive, and a nonadaptive variant. For Adalina, the estimate is

$$\widehat{\boldsymbol{\phi}}^{\text{Adalina}} \coloneqq \frac{1}{T} \sum_{t=1}^{T} (u_{S_t} - \widehat{\lambda}) \cdot \mathbf{z}_{S_t} + \mathbf{c} \;\text{ where }\; \widehat{\lambda} \coloneqq \frac{1}{T} \sum_{t=1}^{T} u_{S_t},$$

whereas the estimate produced by the nonadaptive KernelSHAP is

$$\widehat{\boldsymbol{\phi}}^{\text{Nonadaptive}} \coloneqq \frac{1}{T} \sum_{t=1}^{T} (u_{S_t} - \alpha \cdot s_t) \cdot \mathbf{z}_{S_t} + \mathbf{c} \;\text{ where }\; \alpha \coloneqq \frac{u_{[n]} - u_\emptyset}{n}. \tag{7}$$

As shown in Figure 1a, neither consistently outperforms the other. As we will show in the next section, there are infinitely many *provably* adaptive randomized algorithms with linear memory and linear query complexity, but this lack of consistent dominance persists. This calls into question whether we can determine in advance which algorithm is best suited to the characteristic function U at hand.



(a) None consistently outperforms the other.

(b) Rank correlation between $\widehat{\beta}$ and $\widehat{\text{MSE}}$.

Figure 1: All curves are averaged over 10 random seeds, with standard deviations reported. Only the final estimates are used. In the right plot, $\widehat{\beta}$ is defined in Eq. (11), and $\widehat{\text{MSE}} \coloneqq \|\widehat{\phi} - \phi\|_2^2$.

4 Main Results

4.1 Theoretical Foundation

We begin by observing that, if a characteristic function V satisfies $V(S) = f(s)$ for some function $f\colon [n-1] \to \mathbb{R}$, then

$$\varphi(V) = \mathbf{0}_n,$$

where φ denotes the mapping $U \mapsto \boldsymbol{\varphi}$ defined in Eq. (2). We refer to f as a *control variate*. Indeed, for every $i \in [n]$,

$$\varphi_i(V) = \sum_{s=1}^{n-1} \binom{n-1}{s-1} p_s \cdot f(s) - \sum_{s=1}^{n-1} \binom{n-1}{s} p_{s+1} \cdot f(s) = \sum_{s=1}^{n-1} \left(\frac{1}{n} - \frac{1}{n}\right) \cdot f(s) = 0.$$

As a result, for any fixed $\lambda_f \in \mathbb{R}$,

$$\widehat{\boldsymbol{\phi}}^{f,\text{unbiased}} \coloneqq \frac{1}{T} \sum_{t=1}^{T} [u_{S_t} - \lambda_f f(s_t)] \cdot \mathbf{z}_{S_t} + \mathbf{c}.$$

is an unbiased estimator of $\boldsymbol{\phi}$. This provides a more general perspective on why the nonadaptive KernelSHAP in Eq. (7) works. Under the property in Eq. (6), which is provided by the optimal choice of the sampling distribution $\mathbf{q}$, the MSE is

$$\text{MSE}_{f,\text{unbiased}} = \frac{\mathbb{E}[\|u_{\mathbf{S}} - \lambda_f f(\mathbf{s}) \cdot \mathbf{z}_{\mathbf{S}}\|_2^2] - \|\boldsymbol{\varphi}\|_2^2}{T} = \frac{nD \cdot \mathbb{E}[(u_{\mathbf{S}} - \lambda_f f(\mathbf{s}))^2] - \|\boldsymbol{\varphi}\|_2^2}{T},$$

Setting the derivative of $\text{MSE}_{f,\text{unbiased}}$ with respect to λ_f equal to zero gives the optimal coefficient that minimizes it,

$$\lambda_f^* = \frac{\mathbb{E}[f(\mathbf{s}) u_{\mathbf{S}}]}{\mathbb{E}[f(\mathbf{s})^2]}. \tag{8}$$

Although $\mathbb{E}[f(\mathbf{s})^2]$ can be computed in $O(n)$ time, $\mathbb{E}[f(\mathbf{s}) u_{\mathbf{S}}]$ is generally intractable to compute in advance and must therefore be estimated. This motivates the following prototype adaptive randomized estimator,

$$\widehat{\boldsymbol{\phi}}^f \coloneqq \widehat{\boldsymbol{\varphi}}^f + \mathbf{c} \text{ where } \widehat{\varphi}^f \coloneqq \frac{1}{T} \sum_{t=1}^{T} [u_{S_t} - \widehat{\lambda}_f f(s_t)] \cdot \mathbf{z}_{S_t} \text{ and } \widehat{\lambda}_f \coloneqq \frac{1}{T} \sum_{t=1}^{T} \frac{u_{S_t} f(s_t)}{\mathbb{E}[f(\mathbf{s})^2]}. \tag{9}$$

This formulation includes Adalina as the special case $f \equiv 1$. Although $\widehat{\boldsymbol{\phi}}^f$ is biased, Eq. (6) enables us to establish that its MSE converges to the MSE of the estimator obtained using the optimal coefficient. Specifically,

$$\widehat{\boldsymbol{\phi}}^{f*} \coloneqq \widehat{\boldsymbol{\varphi}}^{f*} + \mathbf{c} \text{ where } \widehat{\boldsymbol{\varphi}}^{f*} \coloneqq \frac{1}{T} \sum_{t=1}^{T} [u_{S_t} - \lambda_f^* f(s_t)] \cdot \mathbf{z}_{S_t}. \tag{10}$$

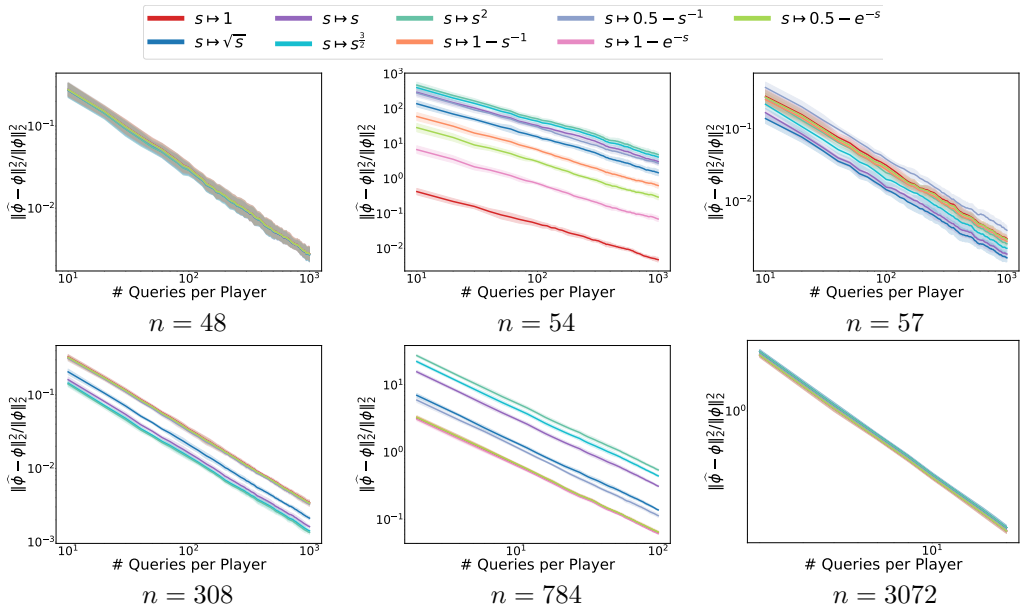


Figure 2: All curves are averaged over 10 random seeds, with standard deviations reported. Each curve corresponds to a particular choice of f, with the estimate produced according to Eq. (9).

Theorem 1 (Adaptivity). *For $T \geq 2$, we have*

$$\mathbb{E}[\|\widehat{\phi}^f - \phi\|_2^2] \leq \mathbb{E}[\|\widehat{\phi}^{f*} - \phi\|_2^2] + \frac{15nDQ_f^2\|\mathbf{u}\|_\infty^2}{T^2} \quad \textit{where } Q_f := \frac{\max_{s\in[n-1]} f(s)^2}{\mathbb{E}[f(\mathbf{s})^2]}.$$

This result provides an exact extension of the corresponding guarantee for Adalina (Li et al., 2026a, Theorem 4.1). We comment that their proof cannot be directly reused, since it relies on the identity $\mathbb{E}[\widehat{\varphi}^f] = \left(1 - \frac{1}{T}\right)\varphi$ provided that $f \equiv 1$. In general, we instead have $\mathbb{E}[\widehat{\varphi}^f] = \varphi - \frac{1}{T}\frac{\mathbb{E}[u_{\mathbf{S}} f(\mathbf{s})^2 \cdot \mathbf{z}_{\mathbf{S}}]}{\mathbb{E}[f(\mathbf{s})^2]}$, which requires a different decomposition argument.

Intuitively, if accurately estimating λ_f^* requires more queries than directly estimating ϕ using Eq. (4), then the query complexity of the estimator in Eq. (9) may increase. This motivates identifying a quantity that characterizes when the linear query complexity is preserved. The quantity Q_f, defined in Theorem 1, serves this purpose.

Theorem 2 (Linear Query Complexity). *If $Q_f \in \Theta(1)$ and $\|\mathbf{u}\|_\infty \leq C$ as $n \to \infty$, it requires $T \geq \frac{30nDQ^2C^2}{\epsilon^2} \log\frac{4}{\delta}$ queries to achieve*

$$P(\|\widehat{\phi}^f - \phi\|_2 \geq \epsilon) \leq \delta.$$

In particular, $Q_f \in \Theta(1)$ if and only if $Q_f \in O(1)$.

Equipped with Theorem 2, we are ready to introduce a class of *provably* adaptive algorithms by identify a sufficient condition on the control variate f that ensures $Q_f \in \Theta(1)$.

Proposition 1 (Choice of Control Variates). *We have $Q_f \in \Theta(1)$ as $n \to \infty$ if either (i) there exist two constants L and U such that $0 < L \leq |f(s)| \leq U$ eventually or (ii) $f(s) = s^\gamma$ with $\gamma \geq 0$.*

Using Proposition 1, we first observe that setting $f(s) = s$ in Eq. (9) turns the nonadaptive KernelSHAP into an adaptive estimator while preserving linear query complexity. Moreover, we can construct many theoretically sound control variates, including the mapping $s \mapsto 1 - e^{-s}$. In data attribution, where $U(S)$ represents the accuracy of a model trained on S, $U(S)$ is typically expected to remain bounded as $s \to \infty$. In this setting, $s \mapsto 1 - e^{-s}$ may be better suited to U than $s \mapsto s^\gamma$, where $\gamma > 0$, in terms of minimizing the MSE. In summary, Proposition 1 provides guidance for designing theoretically justified control variates that match the structure of the characteristic function

U at hand. As confirmed in Figure 2, different choices of f are better suited to different characteristic functions.

Next, we demonstrate that the best-performing control variate can be selected online without access to $\boldsymbol{\phi}$. Since the MSE measures the expected squared approximation error, we use it as the selection criterion. Based on Theorem 1, it suffices to consider the MSE of the optimal counterpart $\widehat{\boldsymbol{\phi}}^{f*}$, defined in Eq. (10). By Eq. (6),

$$\mathrm{MSE}_{f*} \coloneqq \mathbb{E}[\|\widehat{\boldsymbol{\phi}}^{f*} - \boldsymbol{\phi}\|_2^2] \propto \mathbb{E}[(u_{\mathbf{S}} - \lambda_f^* f(\mathbf{s}))^2].$$

Using Eq. (8), we obtain $\mathbb{E}[(u_{\mathbf{S}} - \lambda_f^* f(\mathbf{s}))^2] = \mathbb{E}[u_{\mathbf{S}}^2] - (\lambda_f^*)^2 \cdot \mathbb{E}[f(\mathbf{s})^2]$. The first term is independent of f. Thus, minimizing the MSE amounts to maximizing

$$\beta_f \coloneqq (\lambda_f^*)^2 \cdot \mathbb{E}[f(\mathbf{s})^2].$$

This suggests estimating λ_f^*, which already appears when computing $\widehat{\boldsymbol{\phi}}^f$ in Eq. (9). We therefore estimate β_f by

$$\widehat{\beta}_f \coloneqq \widehat{\lambda}_f^2 \cdot \mathbb{E}[f(\mathbf{s})^2] \;\text{ where }\; \widehat{\lambda}_f = \frac{1}{T}\sum_t^T \frac{u_{S_t} f(s_t)}{\mathbb{E}[f(\mathbf{s})^2]}. \tag{11}$$

In other words, $\widehat{\beta}_f$ can be computed without increasing the memory or query complexity. Moreover, selection based on $\widehat{\beta}_f$ enjoys a theoretical guarantee.

Theorem 3 (Selection Guarantee)**.** *Given that* $\Delta \coloneqq \beta_{f_1} - \beta_{f_2}$*, we have*

$$P((\widehat{\beta}_{f_1} - \widehat{\beta}_{f_2}) \cdot \Delta \le 0) \le 6 \exp\left(-\frac{T\Delta^2}{32\|\mathbf{u}\|_\infty^4 Q_{\max}^2}\right) \textit{where}\;\; Q_{\max} \coloneqq \max(Q_{f_1}, Q_{f_2}).$$

Accordingly, the better estimator between $\widehat{\boldsymbol{\phi}}^{f_1}$ and $\widehat{\boldsymbol{\phi}}^{f_2}$ is selected correctly with probability at least $1-\delta$ whenever

$$T \ge \frac{32\|\mathbf{u}\|_\infty^4 Q_{\max}^2}{\Delta^2} \log\frac{6}{\delta}.$$

This suggests that the accuracy of the selection depends primarily on $|\Delta| \propto |\mathrm{MSE}_{f_1*} - \mathrm{MSE}_{f_2*}|$.

To verify this theoretical result, we considered the control variates $\mathcal{F}$ presented in Figure 2. After collecting a specified number of subsets, we obtained $\{\widehat{\beta}_f\}_{f\in\mathcal{F}}$ and $\{\|\widehat{\boldsymbol{\phi}}^f - \boldsymbol{\phi}\|_2^2\}_{f\in\mathcal{F}}$. We then computed the Kendall rank correlation coefficient τ and Spearman rank correlation coefficient ρ between these two quantities. The results are presented in Figure 1b. A noteworthy case occurs when $n = 48$. The correlation coefficients are significantly low in Figure 1b, while the approximation-error curves for the different control variates are nearly indistinguishable in Figure 2. This observation supports our claim that selection accuracy mainly depends on $|\Delta|$. In other words, the low correlation coefficients occur when the performances of the different control variates are nearly identical, in which case selecting one at random is already satisfactory.

4.2 BestSHAP

Building on the aforementioned theoretical foundation, we introduce BestSHAP, a randomized algorithm capable of selecting the best among many adaptive estimators. Its procedure is summarized in Algorithm 1. The algorithm samples a single sequence of subsets and uses it to compute all the estimates required to select the best output $\widehat{\boldsymbol{\phi}}$ for a given characteristic function U. Since $\mathcal{F}$ is fixed and $Q_f \in \Theta(1)$ for every $f \in \mathcal{F}$, BestSHAP has linear memory and query complexities.

We next verify that BestSHAP is the best-performing variant among existing linear-memory KernelSHAP methods. Following Chen et al. (2025), all existing variants except Adalina can be expressed using Eq. (7), where $\alpha \in \{0, \frac{u_{[n]}-u_\emptyset}{n}\}$ and $q_s \propto g(s)$ with $g \in \{s \mapsto \frac{1}{s(n-s)}, s \mapsto 1, s \mapsto \frac{1}{\sqrt{s(n-s)}}\}$. These choices of $\mathbf{q}$ correspond to kernel sampling (Lundberg & Lee, 2017), leverage sampling (Musco & Witter, 2025), and modified sampling (Chen et al., 2025), respectively. We also consider SHAP-IQ (Fumagalli et al., 2023) as a variant, as it sets $q_s \propto \frac{1}{s(n-s)}$ and $\lambda_f f(s_t) \equiv u_\emptyset$ in Eq. 4.1. As shown in Figure 3, BestSHAP outperforms all the other variants. Notably, BestSHAP can incorporate any adaptive variant and select the best-performing one.

Algorithm 1 BestSHAP (**Best**-of-Many Adaptive Approximations for the **Shap**ley Value)

Input: Total number of samples T, a candidate set of control variates $\mathcal{F}$, and a vector of characteristic functions $\mathbf{u}$
Output: Estimate $\widehat{\boldsymbol{\phi}}$

$A_f \leftarrow \mathbb{E}[f(\mathbf{s})^2]$ for every $f \in \mathcal{F}$ // Eq. (15)
$\widehat{\boldsymbol{\varphi}} \leftarrow \mathbf{0}_n$
$\widehat{\mathbf{v}}_f \leftarrow \mathbf{0}_n$ and $\widehat{\lambda}_f \leftarrow 0$ for every $f \in \mathcal{F}$
for $t = 1, 2, \ldots, T$ **do**
  Sample a subset size s with probability q_s^* // Eq. (5)
  Sample a subset S of size s uniformly from $\{S \subseteq [n] \mid |S| = s\}$
  $\widehat{\boldsymbol{\varphi}} \leftarrow (1 - \frac{1}{t}) \cdot \widehat{\boldsymbol{\varphi}} + \frac{1}{t} \cdot u_S \cdot \mathbf{z}_S$ // Eq. (3)
  $\widehat{\mathbf{v}}_f \leftarrow (1 - \frac{1}{t}) \cdot \widehat{\mathbf{v}}_f + \frac{1}{t} \cdot f(|S|) \cdot \mathbf{z}_S$ for every $f \in \mathcal{F}$
  $\widehat{\lambda}_f \leftarrow (1 - \frac{1}{t}) \cdot \widehat{\lambda}_f + \frac{1}{t} \cdot \frac{f(|S|) \cdot u_S}{A_f}$ for every $f \in \mathcal{F}$
$\widehat{\beta}_f \leftarrow \widehat{\lambda}_f^2 \cdot A_f$ for every $f \in \mathcal{F}$
$F \leftarrow \arg\max_{f \in \mathcal{F}} \widehat{\beta}_f$
$\widehat{\boldsymbol{\phi}} \leftarrow \widehat{\boldsymbol{\varphi}} - \widehat{\lambda}_F \cdot \widehat{\mathbf{v}}_F + \frac{u_{[n]} - u_\emptyset}{n} \cdot \mathbf{1}_n$

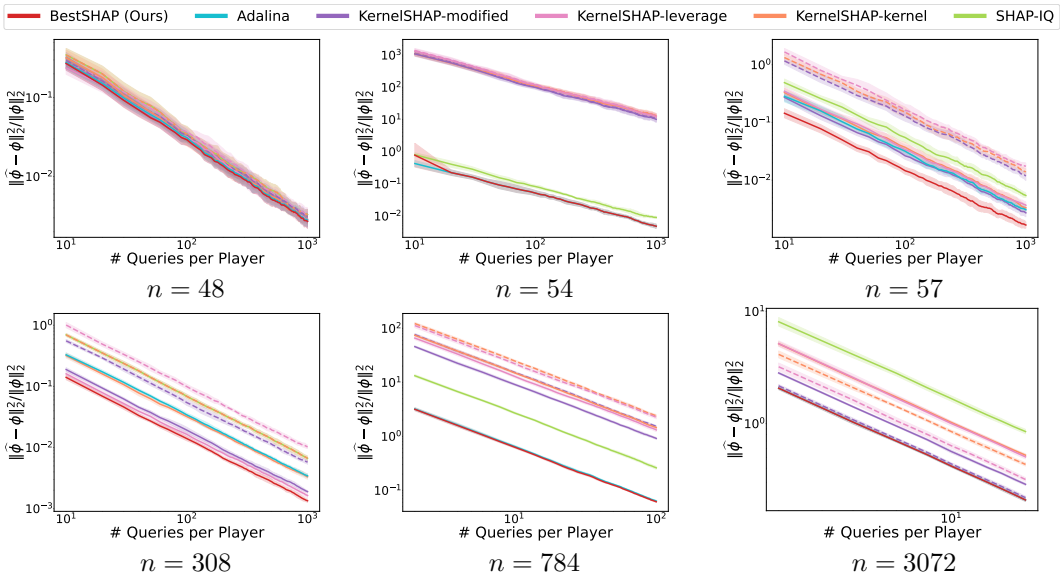


Figure 3: All curves are averaged over 10 random seeds, with standard deviations reported. The dashed lines correspond to the use of $\alpha = 0$ in Eq. (7), whereas the solid lines use $\alpha = \frac{u_{[n]} - u_\emptyset}{n}$.

5 Experiments

In our experiments, we adopt the experimental setting of Li et al. (2026a). More experimental details and results can be found in the Appendix. Our goals are to verify the following: (i) whether the selection criterion $\widehat{\beta}$ used by BestSHAP aligns well with the actual error $\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2^2$, (ii) whether BestSHAP can achieve the best performance among KernelSHAP variants by running many adaptive variants in parallel and selecting the best-performing one, and (iii) how well BestSHAP performs relative to existing linear-memory baselines. For our BestSHAP, $\mathcal{F} := \{s \mapsto 1, s \mapsto \sqrt{s}, s \mapsto s, s \mapsto s^{\frac{3}{2}}, s \mapsto s^2, s \mapsto 1 - s^{-1}, s \mapsto 0.5 - s^{-1}, s \mapsto 1 - e^{-s}, s \mapsto 0.5 - e^{-s}\}$.

Characteristic functions. To evaluate $\widehat{\mathrm{MSE}} := \|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2^2$, we use characteristic functions whose Shapley values can be computed exactly in polynomial time. Specifically, we consider the path-dependent characteristic functions constructed from a trained gradient-boosting decision tree $\mathcal{T}$ and

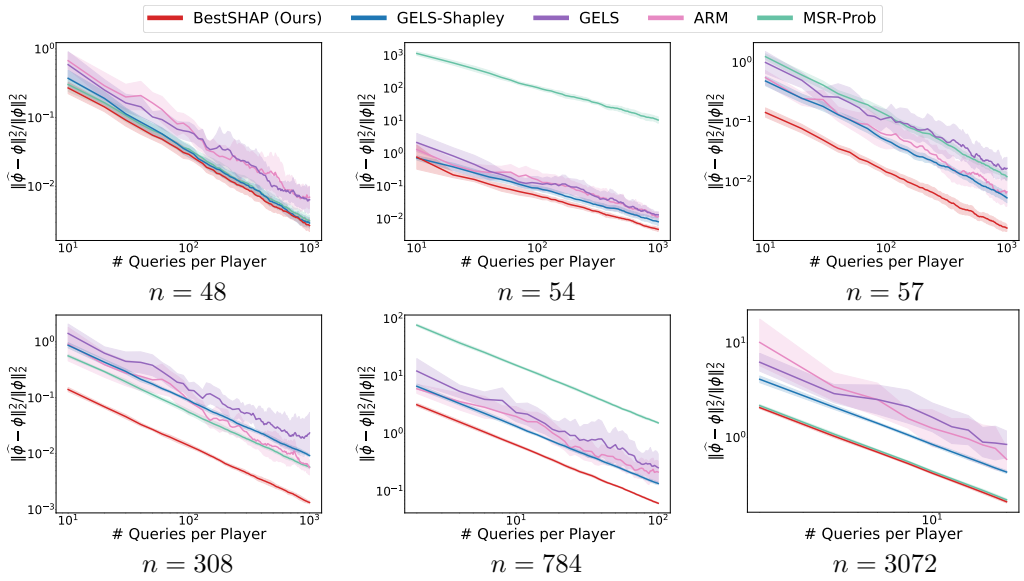


Figure 4: All curves are averaged over 10 random seeds, with standard deviations reported.

an arbitrary instance $\mathbf{x}$ for which $\mathcal{T}$ produces a prediction (Lundberg et al., 2020, Algorithm 1). The Shapley values of these characteristic functions can be computed exactly using TreeGrad-SHAP (Li et al., 2026b, Algorithm 7) in $O(T_{\mathcal{T}}LD)$ time, where $T_{\mathcal{T}}$, D, and L denote the number of trees, the average number of leaf nodes per tree, and the average tree depth, respectively. We use 12 OpenML datasets to train the gradient-boosting trees. The number of players is therefore equal to the number of features in each dataset, yielding $n \in \{32, 48, 50, 51, 54, 57, 77, 81, 308, 561, 784, 3072\}$. The scikit-learn library is employed to train the trees, with $T_{\mathcal{T}} = 10$. For each characteristic function, the estimates are repeated using 10 different random seeds. All other sources of randomness are fixed using the random seed 2026 to ensure reproducibility.

Baselines. We consider all linear-memory KernelSHAP variants (Chen et al., 2025), including SHAP-IQ (Fumagalli et al., 2023) and Adalina (Li et al., 2026a), as well as GELS-Shapley and GELS (Li & Yu, 2024a), ARM (Kolpaczki et al., 2024), and MSR-Prob (Witter et al., 2025).

Results. The empirical correlation between $\widehat{\beta}$ and $\widehat{\mathrm{MSE}}$ is reported in Figure 1b. In most cases, the correlation coefficients exceed 0.8, indicating strong rank alignment. For the cases with relatively low correlation coefficients, namely $n \in \{48, 561\}$, Figures 2 and 5 show that the approximation-error curves for the different control variates are nearly identical. In these cases, random selection makes little practical difference. As shown in Figures 3 and 6, BestSHAP consistently outperforms the other KernelSHAP variants, not to mention that its performance could be further improved by incorporating additional practically suitable control variates. Compared with the other linear-memory baselines, BestSHAP also consistently achieves better performance, as shown in Figures 4 and 7.

6 Conclusion

In conclusion, we introduced BestSHAP, a linear-memory, linear-query-complexity randomized algorithm that runs many adaptive estimators and selects the best-performing one without access to oracle information. BestSHAP is theoretically grounded, with a selection procedure that is highly accurate, and its selection criterion $\widehat{\beta}$ aligns well empirically with the actual squared approximation error $\widehat{\mathrm{MSE}}$. It is also practically flexible, since its candidate set can be expanded to include control variates tailored to different structures of U. Our experiments show that BestSHAP identifies the best-performing variant within the linear-memory KernelSHAP framework and consistently outperforms the other linear-memory baselines. Overall, our work demonstrates that the best-performing estimator can be selected without requiring the ground-truth Shapley values.

AI USE STATEMENT

AI was used solely to improve the clarity and readability of the text. Everything else, including the content, main ideas, proofs, experimental design, and coding, was all human-generated.

REFERENCES

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International workshop on ambient assisted living*, pp. 216–223. Springer, 2012. URL https://doi.org/10.1007/978-3-642-35395-6_30.

James P. Bridge, Sean B. Holden, and Lawrence C. Paulson. Machine learning for first-order theorem proving: Learning to select a good heuristic. *Journal of automated reasoning*, 53:141–172, 2014. URL https://doi.org/10.1007/s10817-014-9301-5.

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. ProxySPEX: Inference-efficient interpretability via sparse feature interactions in LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KI8qan2EA7.

A. Charnes, B. Golany, M. Keane, and J. Rousseau. Extremal principle solutions of games in characteristic function form: Core, Chebychev and Shapley value generalizations. In *Econometrics of planning and efficiency*, pp. 123–133. Springer, 1988. URL https://doi.org/10.1007/978-94-009-3677-5_7.

Tyler Chen, Akshay Seshadri, Mattia Jacopo Villani, Pradeep Niroula, Shouvanik Chakrabarti, Archan Ray, Pranav Deshpande, Romina Yalovetzky, Marco Pistoia, and Niraj Kumar. A unified framework for provably efficient algorithms to estimate shapley values. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KQLCim1NhT.

Ian Covert and Su-In Lee. Improving KernelSHAP: Practical Shapley value estimation using linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3457–3465. PMLR, 2021. URL https://proceedings.mlr.press/v130/covert21a.html.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OPyWRrcjVQw.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=IEMLNF4gK4.

Fabian Fumagalli, Landon Butler, Justin Singh Kang, Kannan Ramchandran, and R. Teal Witter. An odd estimator for Shapley values. In *The Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=xILwgiWAUk.

Fabian Fumagalli, R. Teal Witter, and Christopher Musco. PolySHAP: Extending kernelSHAP with interaction-informed polynomial regression. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=M19J8UGguq.

Michel Grabisch. *Set functions, games and capacities in decision making*, volume 46 of *Theory and Decision Library C*. Springer, 2016. URL https://doi.org/10.1007/978-3-319-30690-2.

Yexiao He, Ziyao Wang, Zheyu Shen, Guoheng Sun, Yucong Dai, Yongkai Wu, Hongyi Wang, and Ang Li. SHED: Shapley-based automated dataset refinement for instruction fine-tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Gqou8PRgWq.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the Shapley value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1167–1176. PMLR, 2019. URL https://proceedings.mlr.press/v89/jia19a.html.

François Kawala, Ahlame Douzal-Chouakria, Eric Gaussier, and Eustache Dimert. Prédictions d'activité dans les réseaux sociaux en ligne. In *Actes de la Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI)*, pp. 16, 2013. URL https://hal.science/hal-00881395.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the Shapley value without marginal contributions. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence*, volume 38, pp. 13246–13255, 2024. URL https://doi.org/10.1609/aaai.v38i12.29225.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Weida Li and Yaoliang Yu. Faster approximation of probabilistic and distributional values via least squares. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=lvSMIsztka.

Weida Li and Yaoliang Yu. One sample fits all: Approximating all probabilistic values simultaneously and efficiently. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=AUg9D2VjcF.

Weida Li, Yaoliang Yu, and Bryan Kian Hsiang Low. Adalina: Adaptive linear approximation for the Shapley value and beyond. In *The Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=hMTCuqjCW8.

Weida Li, Yaoliang Yu, and Bryan Kian Hsiang Low. TreeGrad-Ranker: Feature ranking via $O(L)$-time gradients for decision trees. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=OcMeNbkN13.

Ziqi Liu, Kiljae Lee, Yuan Zhang, and Weijing Tang. First-order efficiency for probabilistic value estimation via a statistical viewpoint, 2026. URL https://arxiv.org/abs/2605.02827.

MingYu Lu, Soham Gadgil, Chris Lin, Chanwoo Kim, and Su-In Lee. SurrogateSHAP: Training-free contributor attribution for text-to-image (T2I) models. In *The Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Zdx1hFJarp.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *The Thirty-first Annual Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67, 2020. URL https://doi.org/10.1038/s42256-019-0138-9.

Renata C. B. Madeo, Clodoaldo A. M. Lima, and Sarajane M. Peres. Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pp. 46–52, 2013. URL https://doi.org/10.1145/2480362.2480373.

Christopher Musco and R. Teal Witter. Provably accurate Shapley value estimation via leverage score sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wg3rBImn3O.

Byron P. Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2–3):577–584, 2005. URL https://doi.org/10.1016/j.nima.2004.12.018.

L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, Volume II*, pp. 307–318. Princeton University Press, 1953. URL https://doi.org/10.1515/9781400881970-018.

Xiao Tian, Jue Fan, Rachael Hwee Ling Sim, Zixuan Wang, Nancy F. Chen, and Bryan Kian Hsiang Low. Is data Shapley not better than random in data selection? Ask NASH. In *The Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vMsrm8UGGC.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data Shapley in one training run. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HD6bWcj87Y.

R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted monte carlo estimators for Shapley values and probabilistic values. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Qabko39AS5.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Jiayao Zhang, Qiheng Sun, Jinfei Liu, Li Xiong, Jian Pei, and Kui Ren. Efficient sampling approaches to Shapley value approximation. *Proceedings of the ACM on Management of Data*, 1 (1):1–24, 2023. URL https://doi.org/10.1145/3588728.

A MISSING PROOFS

Theorem 1 (Adaptivity)**.** *For $T \geq 2$, we have*

$$\mathbb{E}[\|\widehat{\boldsymbol{\phi}}^f - \boldsymbol{\phi}\|_2^2] \leq \mathbb{E}[\|\widehat{\boldsymbol{\phi}}^{f*} - \boldsymbol{\phi}\|_2^2] + \frac{15nDQ_f^2\|\mathbf{u}\|_\infty^2}{T^2} \quad \textit{where } Q_f \coloneqq \frac{\max_{s\in[n-1]} f(s)^2}{\mathbb{E}[f(\mathbf{s})^2]}.$$

Proof. For notational simplicity, we suppress the subscript/superscript f throughout. In addition, we write $A_{ij} \coloneqq \mathbb{E}[f(\mathbf{s})^i u_{\mathbf{S}}^j]$; accordingly, $\lambda^* = \frac{A_{11}}{A_{20}}$. Recall that $\mathbb{E}[f(\mathbf{s}) \cdot \mathbf{z}_{\mathbf{S}}] = \mathbf{0}$.

We start by observing that $\widehat{\boldsymbol{\varphi}}^f$ is slightly biased. Specifically,

$$\begin{aligned}
\widehat{\boldsymbol{\varphi}} &= \frac{1}{T}\sum_{t=1}^{T}[u_{S_t} - \widehat{\lambda} f(s_t)] \cdot \mathbf{z}_{S_t} \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T}[u_{S_t} - \lambda^* f(s_t)] \cdot \mathbf{z}_{S_t}}_{\widehat{\boldsymbol{\varphi}}^*} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}(\lambda^* - \widehat{\lambda}) f(s_t) \cdot \mathbf{z}_{S_t}}_{\widehat{\mathbf{r}}}.
\end{aligned}$$

In particular, $\mathbb{E}[\widehat{\boldsymbol{\varphi}}^*] = \boldsymbol{\varphi}$ and

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbf{r}}] &= \underbrace{\mathbb{E}\left[\frac{\lambda^*}{T}\sum_{t=1}^{T} f(s_t) \cdot \mathbf{z}_{S_t}\right]}_{0} - \mathbb{E}\left[\frac{1}{T^2 A_{20}}\sum_{t=1}^{T} u_{S_t} f(s_t)^2 \cdot \mathbf{z}_{S_t}\right] \\
&\quad - \underbrace{\mathbb{E}\left[\frac{1}{T^2 A_{20}}\sum_{t=1}^{T}\sum_{k\neq t} u_{S_k} f(s_k) f(s_t) \cdot \mathbf{z}_{S_t}\right]}_{0} \\
&= -\frac{\mathbb{E}[u_{\mathbf{S}} f(\mathbf{s})^2 \cdot \mathbf{z}_{\mathbf{S}}]}{T A_{20}} =: \mathbf{r}.
\end{aligned} \tag{12}$$

To bound $\mathbb{E}[\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_2^2]$, our first step is

$$\mathbb{E}[\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_2^2] = \mathbb{E}[\|\widehat{\boldsymbol{\varphi}} - (\boldsymbol{\varphi} + \mathbf{r})\|_2^2] + \|\mathbf{r}\|_2^2.$$

Here,

$$\|\mathbf{r}\|_2^2 = \frac{1}{T^2 A_{20}^2}\|\mathbb{E}[u_{\mathbf{S}} f(\mathbf{s})^2 \cdot \mathbf{z}_{\mathbf{S}}]\|^2 \leq \frac{1}{T^2 A_{20}^2}\mathbb{E}[\|u_{\mathbf{S}} f(\mathbf{s})^2 \cdot \mathbf{z}_{\mathbf{S}}\|^2] \leq \frac{nDA_{42}}{T^2 A_{20}^2} \leq \frac{nDQ^2\|\mathbf{u}\|_\infty^2}{T^2}.$$

The next step is to tackle the remaining part $\mathbb{E}[\|\widehat{\boldsymbol{\varphi}} - (\boldsymbol{\varphi} + \mathbf{r})\|_2^2]$. By expansion,

$$\begin{aligned}
\mathbb{E}[\|\widehat{\boldsymbol{\varphi}} - (\boldsymbol{\varphi} + \mathbf{r})\|_2^2] &= \mathbb{E}[\|(\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}) + (\widehat{\mathbf{r}} - \mathbf{r})\|_2^2] \\
&= \mathbb{E}[\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2^2] + \mathbb{E}[\|\widehat{\mathbf{r}} - \mathbf{r}\|_2^2] + 2\mathbb{E}[\langle \widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}, \widehat{\mathbf{r}} - \mathbf{r}\rangle] \\
&= \mathbb{E}[\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2^2] + \mathbb{E}[\|\widehat{\mathbf{r}} - \mathbf{r}\|_2^2] + 2(\mathbb{E}[\langle \widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] - \langle \boldsymbol{\varphi}, \mathbf{r}\rangle).
\end{aligned}$$

We will derive upper bounds for the two terms $\mathbb{E}[\|\widehat{\mathbf{r}} - \mathbf{r}\|_2^2]$ and $\mathbb{E}[\langle \widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] - \langle \boldsymbol{\varphi}, \mathbf{r}\rangle$, separately.

For the first term, $\mathbb{E}[\|\widehat{\mathbf{r}} - \mathbf{r}\|_2]^2 \leq \mathbb{E}[\|\widehat{\mathbf{r}}\|_2^2]$ and

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mathbf{r}}\|_2^2] &= \frac{1}{T^2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\langle(\lambda^* - \widehat{\lambda}) f(s_t) \cdot \mathbf{z}_{S_t}, (\lambda^* - \widehat{\lambda}) f(s_k) \cdot \mathbf{z}_{S_k}\rangle\right] \\
&= \frac{1}{T^2}\Big(T \cdot \mathbb{E}\left[\langle(\lambda^* - \widehat{\lambda}) f(s_t) \cdot \mathbf{z}_{S_t}, (\lambda^* - \widehat{\lambda}) f(s_t) \cdot \mathbf{z}_{S_t}\rangle\right] \\
&\quad + T(T-1) \cdot \mathbb{E}_{t\neq k}\left[\langle(\lambda^* - \widehat{\lambda}) f(s_t) \cdot \mathbf{z}_{S_t}, (\lambda^* - \widehat{\lambda}) f(s_k) \cdot \mathbf{z}_{S_k}\rangle\right]\Big) \\
&= \frac{1}{T^2}\left(TnD \cdot \underbrace{\mathbb{E}[(\lambda^* - \widehat{\lambda})^2 f(s_t)^2]}_{(11)} + T(T-1) \cdot \underbrace{\mathbb{E}_{t\neq k}[\widehat{\lambda}^2 \langle f(s_t) \cdot \mathbf{z}_{S_t}, f(s_k) \cdot \mathbf{z}_{S_k}\rangle]}_{(12)}\right).
\end{aligned}$$

For the part (11),

$$\mathbb{E}[(\lambda^*)^2 f(\mathbf{s})^2] = \frac{A_{11}^2}{A_{20}},$$

and

$$\begin{aligned}
\mathbb{E}[\lambda^* \widehat{\lambda} f(s_t)^2] &= \frac{A_{11}}{TA_{20}^2}\mathbb{E}\left[\sum_{m=1}^{T} f(s_m)u_{S_m} f(s_t)^2\right] \\
&= \frac{A_{11}}{TA_{20}^2}\left(\mathbb{E}[f(\mathbf{s})^3 u_{\mathbf{S}}] + (T-1)\cdot \mathbb{E}_{m\neq t}[f(s_m)u_{S_m} f(s_t)^2]\right) \\
&= \frac{A_{11}}{TA_{20}^2}\left(A_{31} + (T-1)\cdot A_{11}A_{20}\right).
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
\mathbb{E}[\widehat{\lambda}^2 f(s_t)^2] &= \frac{1}{T^2A_{20}^2}\mathbb{E}\left[\sum_{m=1}^{T}\sum_{l=1}^{T} f(s_m)u_{S_m} f(s_l)u_{S_l} f(s_t)^2\right] \\
&= \frac{1}{T^2A_{20}^2}\left(\mathbb{E}\left[\sum_{l=1}^{T} f(s_t)^3 u_{S_t} f(s_l)u_{S_l}\right]\right. \\
&\quad \left. + (T-1)\cdot \mathbb{E}_{m\neq t}\left[\sum_{l=1}^{T} f(s_m)u_{S_m} f(s_l)u_{S_l} f(s_t)^2\right]\right) \\
&= \frac{1}{T^2A_{20}^2}\{[A_{42} + (T-1)A_{31}A_{11}] \\
&\quad + (T-1)\cdot[A_{11}A_{31} + A_{22}A_{20} + (T-2)A_{11}^2 A_{20}]\}.
\end{aligned}$$

Putting them together yields

$$\begin{aligned}
(11) &= \mathbb{E}[(\lambda^*)^2 f(\mathbf{s})^2] + \mathbb{E}[\widehat{\lambda}^2 f(s_t)^2] - 2\mathbb{E}[\lambda^* \widehat{\lambda} f(s_t)^2] \\
&= \frac{A_{42}}{T^2A_{20}^2} + \frac{(T-1)A_{22}A_{20}}{T^2A_{20}^2} - \frac{2A_{31}A_{11}}{T^2A_{20}^2} - \frac{(T-2)A_{20}A_{11}^2}{T^2A_{20}^2} \\
&\leq \frac{Q^2\|\mathbf{u}\|_\infty^2}{T} + \frac{Q^2\|\mathbf{u}\|_\infty^2}{T} + \frac{2Q^2\|\mathbf{u}\|_\infty^2}{T} + 0 = \frac{4Q^2\|\mathbf{u}\|_\infty^2}{T}
\end{aligned}$$

For the part (12),

$$\begin{aligned}
(12) &= \frac{1}{T^2A_{20}^2}\mathbb{E}_{t\neq k}\left[\sum_{m=1}^{T}\sum_{l=1}^{T} f(s_m)u_{S_m} f(s_l)u_{S_l}\langle f(s_t)\cdot \mathbf{z}_{S_t}, f(s_k)\cdot \mathbf{z}_{S_k}\rangle\right] \\
&= \frac{1}{T^2A_{20}^2}\left(2\cdot \mathbb{E}_{t\neq k}\left[\sum_{l=1}^{T} f(s_l)u_{S_l}\langle f(s_t)^2 u_{S_t}\cdot \mathbf{z}_{S_t}, f(s_k)\cdot \mathbf{z}_{S_k}\rangle\right]\right. \\
&\quad \left. + (T-2)\cdot \mathbb{E}_{\substack{t\neq k\\ m\neq t\\ m\neq k}}\left[\sum_{l=1}^{T} f(s_m)u_{S_m} f(s_l)u_{S_l}\langle f(s_t)\cdot \mathbf{z}_{S_t}, f(s_k)\cdot \mathbf{z}_{S_k}\rangle\right]\right) \\
&= \frac{1}{T^2A_{20}^2}[2\cdot \|\mathbb{E}[f(\mathbf{s})^2 u_{\mathbf{S}}\cdot \mathbf{z}_{\mathbf{S}}]\|_2^2 + (T-2)\cdot 0] \\
&\leq \frac{2}{T^2A_{20}^2}\mathbb{E}[\|f(\mathbf{s})^2 u_{\mathbf{S}}\cdot \mathbf{z}_{\mathbf{S}}\|_2^2] = \frac{2nDA_{42}}{T^2A_{20}^2} \leq \frac{2nDQ^2\|\mathbf{u}\|_\infty^2}{T^2}.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mathbf{r}} - \mathbf{r}\|_2^2] \leq \mathbb{E}[\|\widehat{\mathbf{r}}\|_2^2] &\leq \frac{1}{T^2}\left[TnD\cdot \frac{4Q^2\|\mathbf{u}\|_\infty^2}{T} + T(T-1)\cdot \frac{2nDQ^2\|\mathbf{u}\|_\infty^2}{T^2}\right] \\
&\leq \frac{6nDQ^2\|\mathbf{u}\|_\infty^2}{T^2}
\end{aligned}$$

What remains is to bound the second term $\mathbb{E}[\langle\widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] - \langle\boldsymbol{\varphi}, \mathbf{r}\rangle$. Observe that,

$$\begin{aligned}\mathbb{E}[\langle\widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] &= \frac{1}{T^2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\langle[u_{S_t} - \lambda^* f(s_t)]\cdot\mathbf{z}_{S_t}, (\lambda^* - \widehat{\lambda})f(s_k)\cdot\mathbf{z}_{S_k}\rangle\right] \\ &= \frac{1}{T^2}\left(T\cdot\underbrace{\mathbb{E}\left[\langle[u_{S_t} - \lambda^* f(s_t)]\cdot\mathbf{z}_{S_t}, (\lambda^* - \widehat{\lambda})f(s_t)\cdot\mathbf{z}_{S_t}\rangle\right]}_{\textcircled{21}}\right. \\ &\quad \left. + T(T-1)\cdot\underbrace{\mathbb{E}_{t\neq k}\left[\langle[u_{S_t} - \lambda^* f(s_t)]\cdot\mathbf{z}_{S_t}, (\lambda^* - \widehat{\lambda})f(s_k)\cdot\mathbf{z}_{S_k}\rangle\right]}_{\textcircled{22}}\right)\end{aligned}$$

Since

$$\mathbb{E}[[u_{\mathbf{S}} - \lambda^* f(\mathbf{s})]f(\mathbf{s})] = A_{11} - \frac{A_{11}}{A_{20}}A_{20} = 0,$$

there is

$$\begin{aligned}\textcircled{21} &= nD\cdot\mathbb{E}[[u_{S_t} - \lambda^* f(s_t)](\lambda^* - \widehat{\lambda})f(s_t)] \\ &= -nD\cdot\mathbb{E}[[u_{S_t} - \lambda^* f(s_t)]\widehat{\lambda}f(s_t)] \\ &= -\frac{nD}{TA_{20}}\mathbb{E}\left[\sum_{k=1}^{T} f(s_k)u_{S_k}[u_{S_t} - \lambda^* f(s_t)]f(s_t)\right] \\ &= -\frac{nD}{TA_{20}}\mathbb{E}[f(\mathbf{s})^2 u_{\mathbf{S}}[u_{\mathbf{S}} - \lambda^* f(\mathbf{s})]] \\ &= \frac{nD}{TA_{20}}\left(\frac{A_{11}A_{31}}{A_{20}} - A_{22}\right) \leq \frac{2nDQ^2\|\mathbf{u}\|_\infty^2}{T}\end{aligned}$$

Since $\mathbb{E}[f(\mathbf{s})\cdot u_{\mathbf{S}}] = \mathbf{0}$, we have

$$\begin{aligned}\textcircled{22} &= -\mathbb{E}_{t\neq k}\left[\langle[u_{S_t} - \lambda^* f(s_t)]\cdot\mathbf{z}_{S_t}, \widehat{\lambda}f(s_k)\cdot\mathbf{z}_{S_k}\rangle\right] \\ &= -\frac{1}{TA_{20}}\mathbb{E}_{t\neq k}\left[\sum_{m=1}^{T}\langle[u_{S_t} - \lambda^* f(s_t)]\cdot\mathbf{z}_{S_t}, f(s_m)u_{S_m}f(s_k)\cdot\mathbf{z}_{S_k}\rangle\right] \\ &= -\frac{1}{TA_{20}}\mathbb{E}_{t\neq k}[\langle[u_{S_t} - \lambda^* f(s_t)]\cdot\mathbf{z}_{S_t}, f(s_k)^2 u_{S_k}\cdot\mathbf{z}_{S_k}\rangle].\end{aligned}$$

According to Eq. (12), $\mathbb{E}[u_{\mathbf{S}}f(\mathbf{s})^2\cdot\mathbf{z}_{\mathbf{S}}] = -TA_{20}\cdot\mathbf{r}$, and thus

$$\textcircled{22} = \langle\boldsymbol{\varphi}, \mathbf{r}\rangle.$$

As a result,

$$\begin{aligned}\mathbb{E}[\langle\widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] - \langle\boldsymbol{\varphi}, \mathbf{r}\rangle &\leq \frac{1}{T^2}\left(T\cdot\frac{2nDQ^2\|\mathbf{u}\|_\infty^2}{T} + T(T-1)\cdot\langle\boldsymbol{\varphi}, \mathbf{r}\rangle\right) - \langle\boldsymbol{\varphi}, \mathbf{r}\rangle \\ &= \frac{2nDQ^2\|\mathbf{u}\|_\infty^2}{T^2} - \frac{1}{T}\langle\boldsymbol{\varphi}, \mathbf{r}\rangle \\ &\leq \frac{2nDQ^2\|\mathbf{u}\|_\infty^2}{T^2} + \frac{1}{T}\|\boldsymbol{\varphi}\|_2\|\mathbf{r}\|_2\end{aligned}$$

By Eq. (12),

$$\|\mathbf{r}\|_2 \leq \frac{1}{TA_{20}}\mathbb{E}[\|u_{\mathbf{S}}f(\mathbf{s})^2\cdot\mathbf{z}_{\mathbf{S}}\|_2] \leq \frac{1}{T}\sqrt{nDQ^2\|\mathbf{u}\|_\infty^2}.$$

On the other hand,

$$|u_{S_t} - \lambda^* f(s_t)| \le \|\mathbf{u}\|_\infty + \frac{|A_{11}| \cdot \max_{s\in[n-1]} |f(s)|}{A_{20}} = \frac{A_{20}\|\mathbf{u}\|_\infty + |A_{11}| \cdot \max_{s\in[n-1]} |f(s)|}{A_{20}} \le 2Q\|\mathbf{u}\|_\infty, \tag{13}$$

which leads to

$$\|\boldsymbol{\varphi}\|_2 \le \mathbb{E}[\|[u_{\mathbf{S}} - \lambda^* f(s_t)] \cdot \mathbf{z}_{\mathbf{S}}\|_2] \le \sqrt{4nDQ^2\|\mathbf{u}\|_\infty^2}$$

Therefore,

$$\mathbb{E}[\langle \widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] - \langle \boldsymbol{\varphi}, \mathbf{r}\rangle \le \frac{4nDQ^2\|\mathbf{u}\|_\infty^2}{T^2}.$$

Finally, we have

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_2^2] &= \mathbb{E}[\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2^2] + \mathbb{E}[\|\widehat{\mathbf{r}} - \mathbf{r}\|_2^2] + 2(\mathbb{E}[\langle \widehat{\boldsymbol{\varphi}}^*, \widehat{\mathbf{r}}\rangle] - \langle \boldsymbol{\varphi}, \mathbf{r}\rangle) + \|\mathbf{r}\|_2^2 \\ &\le \mathbb{E}[\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2^2] + \frac{6nDQ^2\|\mathbf{u}\|_\infty^2}{T^2} + \frac{8nDQ^2\|\mathbf{u}\|_\infty^2}{T^2} + \frac{nDQ^2\|\mathbf{u}\|_\infty^2}{T^2} \\ &= \mathbb{E}[\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2^2] + \frac{15nDQ^2\|\mathbf{u}\|_\infty^2}{T^2}\end{aligned}$$

□

Theorem 4 (Vector concentration inequality (Li et al., 2026a)). *Suppose $\{\mathbf{X}_i\}_{i=1}^M$ are i.i.d. zero-mean random vectors such that $\mathbb{E}[\|\mathbf{X}_i\|_2^2] \le \sigma^2$ and $\|\mathbf{X}_i\|_2 \le C$ almost surely. Then, for every $0 < \epsilon \le \frac{3\sigma^2}{C}$, there is*

$$\mathbb{P}\left(\left\|\frac{1}{M}\sum_{i=1}^M \mathbf{X}_i\right\|_2 \ge \epsilon\right) \le 2\exp\left(-\frac{M\epsilon^2}{4\sigma^2}\right).$$

Theorem 2 (Linear Query Complexity). *If $Q_f \in \Theta(1)$ and $\|\mathbf{u}\|_\infty \le C$ as $n \to \infty$, it requires $T \ge \frac{30nDQ^2C^2}{\epsilon^2}\log\frac{4}{\delta}$ queries to achieve*

$$P(\|\widehat{\boldsymbol{\phi}}^f - \boldsymbol{\phi}\|_2 \ge \epsilon) \le \delta.$$

In particular, $Q_f \in \Theta(1)$ if and only if $Q_f \in O(1)$.

Proof. For notational simplicity, we suppress the subscript/superscript f throughout. Plus, we denote $\max_{s\in[n-1]} |f(s)|$ by F. Recall that $Q = \frac{F^2}{A_{20}}$.

By definition,

$$\begin{aligned}\mathbb{E}[f(\mathbf{s})^2] &= \frac{1}{\sqrt{D}}\sum_{s=1}^{n-1}\frac{f(s)^2}{\sqrt{s(n-s)}} \le \max_{s\in[n-1]} f(s)^2 \cdot \sqrt{D}\sum_{s=1}^{n-1}\frac{1}{\sqrt{s(n-s)}} \\ &= D\max_{s\in[n-1]} f(s)^2.\end{aligned}$$

Since $D \in \Theta(1)$ as $n \to \infty$, we have $Q \in \Omega(1)$. Consequently, $Q \in O(1)$ if and only if $Q \in \Theta(1)$.

Next, notice that

$$\left|\frac{f(s_t)u_{S_t}}{A_{20}}\right| \le \frac{FC}{A_{20}}.$$

Using Hoeffding's inequality, we have

$$P(|\widehat{\lambda} - \lambda^*| \ge \epsilon) \le 2\exp\left(-\frac{TA_{20}^2\epsilon^2}{2F^2C^2}\right).$$

In other words, with probability at least $1 - \frac{\delta}{2}$, there is

$$|\widehat{\lambda} - \lambda^*| < \sqrt{\frac{2F^2C^2}{TA_{20}^2}\log\frac{4}{\delta}}. \tag{14}$$

Then, we apply Theorem 4 to obtain a bound on $\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2$. By Eq. (13),

$$\begin{aligned}\|[u_{S_t} - \lambda^* f(s_t)] \cdot \mathbf{z}_{S_t} - \boldsymbol{\varphi}\|_2 &\leq \|[u_{S_t} - \lambda^* f(s_t)] \cdot \mathbf{z}_{S_t}\|_2 + \|\boldsymbol{\varphi}\|_2 \\ &\leq 2QC\sqrt{nD} + \mathbb{E}[\|[u_{S_t} - \lambda^* f(s_t)] \cdot \mathbf{z}_{S_t}\|_2] \\ &= 4QC\sqrt{nD}\end{aligned}$$

and

$$\mathbb{E}[\|[u_{S_t} - \lambda^* f(s_t)] \cdot \mathbf{z}_{S_t} - \boldsymbol{\varphi}\|_2^2] \leq \mathbb{E}[\|[u_{S_t} - \lambda^* f(s_t)] \cdot \mathbf{z}_{S_t}\|_2^2] \leq 4nDQ^2C^2.$$

Accordingly,

$$P(\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2 \geq \epsilon) \leq 2\exp\left(-\frac{T\epsilon^2}{16nDQ^2C^2}\right)$$

where $\epsilon \leq 3QC\sqrt{nD}$. Then, with probability at least $1 - \frac{\delta}{2}$,

$$\|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2 < \sqrt{\frac{16nDQ^2C^2}{T}\log\frac{4}{\delta}}.$$

Solving $\sqrt{\frac{16nDQ^2C^2}{T}\log\frac{4}{\delta}} \leq 3QC\sqrt{nD}$ yields the condition $T \geq \frac{16}{9}\log\frac{4}{\delta}$. Proceeding, with probability at least $1 - \delta$,

$$\begin{aligned}\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_2 &\leq \|\widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\varphi}}^*\|_2 + \|\widehat{\boldsymbol{\varphi}}^* - \boldsymbol{\varphi}\|_2 \\ &< \frac{1}{T}\sum_{t=1}^{T}\|(\lambda^* - \widehat{\lambda})f(s_t) \cdot \mathbf{z}_{S_t}\|_2 + \sqrt{\frac{16nDQ^2C^2}{T}\log\frac{4}{\delta}} \\ &< \sqrt{\frac{2nDQ^2C^2}{T}\log\frac{4}{\delta}} + \sqrt{\frac{16nDQ^2C^2}{T}\log\frac{4}{\delta}} \\ &< \sqrt{\frac{30nDQ^2C^2}{T}\log\frac{4}{\delta}}\end{aligned}$$

Finally, solving $\sqrt{\frac{30nDQ^2C^2}{T}\log\frac{4}{\delta}} \geq \epsilon$ leads to the conclusion that it requires

$$T \geq \frac{30nDQ^2C^2}{\epsilon^2}\log\frac{4}{\delta}$$

to achieve $P(\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_2 \geq \epsilon) \leq \delta$. In particular, setting $\frac{30nDQ^2C^2}{\epsilon^2}\log\frac{4}{\delta} \geq \frac{16}{9}\log\frac{4}{\delta}$ gives the condition on ϵ that $\epsilon \leq 4QC\sqrt{nD}$. Since $Q, C, D \in \Theta(1)$, this condition can be ignored while deriving the asymptotic query complexity. □

Proposition 1 (Choice of Control Variates). *We have $Q_f \in \Theta(1)$ as $n \to \infty$ if either (i) there exist two constants L and U such that $0 < L \leq |f(s)| \leq U$ eventually or (ii) $f(s) = s^\gamma$ with $\gamma \geq 0$.*

Proof. Recall that

$$\begin{gathered}\mathbb{E}[f(\mathbf{s})^2] = \frac{1}{A}\sum_{s=1}^{n-1}\frac{f(s)^2}{\sqrt{s(n-s)}} \quad \text{where} \quad A = \sum_{s=1}^{n-1}\frac{1}{\sqrt{s(n-s)}}, \\ \text{and} \quad Q_f = \frac{\max_{s\in[n-1]} f(s)^2}{\mathbb{E}[f(\mathbf{s})^2]}.\end{gathered} \tag{15}$$

In particular, by setting $x_s = \frac{s}{n}$,

$$A = \frac{1}{n}\sum_{s=1}^{n-1}\frac{1}{\sqrt{x(1-x)}} \to \int_0^1 \frac{1}{\sqrt{t(1-t)}}\,\mathrm{d}t = \pi.$$

Therefore, if there exist two constants L and U such that $0 < L \leq |f(s)| \leq U$ eventually, we immediately have $\max_{s\in[n-1]} f(s)^2 \in \Theta(1)$ and $\mathbb{E}[f(\mathbf{s})^2] \in \Theta(1)$.

Next, suppose $f(s) = s^\gamma$ with $\gamma \geq 0$. Since $\max_{s\in[n-1]} f(s)^2 = n^{2\gamma}$, suffice it to prove that $\mathbb{E}[f(\mathbf{s})^2] \in \Theta(n^{2\gamma})$. Setting $x_s = \frac{s}{n}$, we have

$$\sum_{s=1}^{n-1} \frac{f(s)^2}{\sqrt{s(n-s)}} = n^{2\gamma} \cdot \frac{1}{n} \sum_{s=1}^{n-1} x^{2\gamma-\frac{1}{2}}(1-x)^{-\frac{1}{2}} \approx n^{2\gamma} \int_0^1 t^{2\gamma-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,\mathrm{d}t$$
$$= n^{2\gamma} \cdot B\left(2\gamma+\frac{1}{2}, \frac{1}{2}\right).$$

Specifically,

$$B\left(2\gamma+\frac{1}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(2\gamma+\frac{1}{2}\right)}{\Gamma\left(2\gamma+1\right)} \cdot \Gamma\left(\frac{1}{2}\right) = \frac{\prod_{k=1}^{2\gamma}\left(k-\frac{1}{2}\right)}{(2\gamma)!} \cdot \frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma(1)} \cdot \Gamma\left(\frac{1}{2}\right) \in \Theta(1).$$

Since $A \in \Theta(1)$, we obtain $\mathbb{E}[f(\mathbf{s})^2] \in \Theta(n^{2\gamma})$, and thus $Q_f \in \Theta(1)$. □

Theorem 3 (Selection Guarantee)**.** *Given that* $\Delta \coloneqq \beta_{f_1} - \beta_{f_2}$*, we have*

$$P((\widehat{\beta}_{f_1} - \widehat{\beta}_{f_2}) \cdot \Delta \leq 0) \leq 6\exp\left(-\frac{T\Delta^2}{32\|\mathbf{u}\|_\infty^4 Q_{\max}^2}\right) \textit{ where } Q_{\max} \coloneqq \max(Q_{f_1}, Q_{f_2}).$$

Proof. Recall that

$$\widehat{\beta}_{f_i} = \widehat{\lambda}_{f_i}^2 \mathbb{E}[f_i(\mathbf{s})^2] = \frac{\widehat{\lambda}_{f_i}}{T} \sum_{t=1}^{T} u_{S_t} f(s_t) \text{ and } Q_{f_i} = \frac{\max_{s\in[n-1]} f_i(s)^2}{\mathbb{E}[f_i(\mathbf{s})^2]}.$$

We will make use of

$$\widehat{\beta}_{f_i}^* = \frac{\lambda_{f_i}^*}{T} \sum_{t=1}^{T} u_{S_t} f_i(s_t).$$

Notice that $\mathbb{E}[\widehat{\beta}_{f_1}^*] - \mathbb{E}[\widehat{\beta}_{f_2}^*] = \Delta$. To apply Hoeffding's inequality, note that

$$|\lambda_{f_1}^* u_{S_t} f_1(s_t) - \lambda_{f_2}^* u_{S_t} f_2(s_t)| \leq 2\|\mathbf{u}\|_\infty^2 Q_{\max}.$$

Therefore, with probability at least $1 - \frac{\delta}{3}$,

$$|(\widehat{\beta}_{f_1}^* - \widehat{\beta}_{f_2}^*) - \Delta| < \sqrt{\frac{8Q_{\max}^2\|\mathbf{u}\|_\infty^4}{T} \log\frac{6}{\delta}}$$

By Eq. (14), for each $i \in \{1, 2\}$, with probability at least $1 - \frac{\delta}{3}$,

$$|\widehat{\lambda}_{f_i} - \lambda_{f_i}^*| < \sqrt{\frac{2F_{f_i}^2\|\mathbf{u}\|_\infty^2}{TA_{f_i}^2} \log\frac{6}{\delta}}$$

where $F_{f_i} \coloneqq \max_{s\in[n-1]} |f_i(s)|$ and $A_{f_i} = \mathbb{E}[f_i(\mathbf{s})^2]$.

Next,

$$|(\widehat{\beta}_{f_1} - \widehat{\beta}_{f_2}) - \Delta| \leq |(\widehat{\beta}_{f_1} - \widehat{\beta}_{f_2}) - (\widehat{\beta}_{f_1}^* - \widehat{\beta}_{f_2}^*)| + |(\widehat{\beta}_{f_1}^* - \widehat{\beta}_{f_2}^*) - \Delta|.$$

Specifically,

$$|\widehat{\beta}_{f_i} - \widehat{\beta}_{f_i}^*| \leq \|\mathbf{u}\|_\infty F_{f_i} \cdot |\widehat{\lambda}_{f_i} - \widehat{\lambda}_{f_i}^*|$$

Combining these, we obtain, with probability at least $1 - \delta$,

$$|\widehat{\beta}_{f_i} - \widehat{\beta}_{f_i}^*| < \sqrt{\frac{2Q_{\max}^2\|\mathbf{u}\|_\infty^4}{T} \log\frac{6}{\delta}} + \sqrt{\frac{2Q_{\max}^2\|\mathbf{u}\|_\infty^4}{T} \log\frac{6}{\delta}} + \sqrt{\frac{8Q_{\max}^2\|\mathbf{u}\|_\infty^4}{T} \log\frac{6}{\delta}}$$
$$= \sqrt{\frac{32Q_{\max}^2\|\mathbf{u}\|_\infty^4}{T} \log\frac{6}{\delta}}.$$

This is equivalent to

$$P(|(\widehat{\beta}_{f_1} - \widehat{\beta}_{f_2}) - \Delta| \geq \epsilon) \leq 6\exp\left(-\frac{T\epsilon^2}{32Q_{\max}^2\|\mathbf{u}\|_\infty^4}\right),$$

and the conclusion follows. □

Table 1: Summary of the datasets used.

Dataset	#Instances	#Features	Source	Task	#Classes	Depth
GPSP Madeo et al. (2013)	9,873	32	https://openml.org/d/4538	classification	5	10
wave_energy	72,000	48	https://openml.org/d/44975	regression	-	30
MinibooNE Roe et al. (2005)	130,064	50	https://openml.org/d/41150	classification	2	20
FOTP Bridge et al. (2014)	6,118	51	https://openml.org/d/1475	classification	6	20
jannis	83,733	54	https://openml.org/d/41168	classification	4	40
spambase	4,601	57	https://openml.org/d/44	classification	2	15
BT (Kawala et al., 2013)	583,250	77	https://openml.org/d/4549	regression	-	20
superconduct	21,263	81	https://openml.org/d/43174	regression	-	10
philippine	5,832	308	https://openml.org/d/41145	classification	2	15
har (Anguita et al., 2012)	10,299	561	https://openml.org/d/1478	classification	6	20
Fashion-MNIST (Xiao et al., 2017)	70,000	784	https://openml.org/d/40996	classification	10	30
CIFAR10_small (Krizhevsky, 2009)	20,000	3,072	https://openml.org/d/40926	classification	10	20

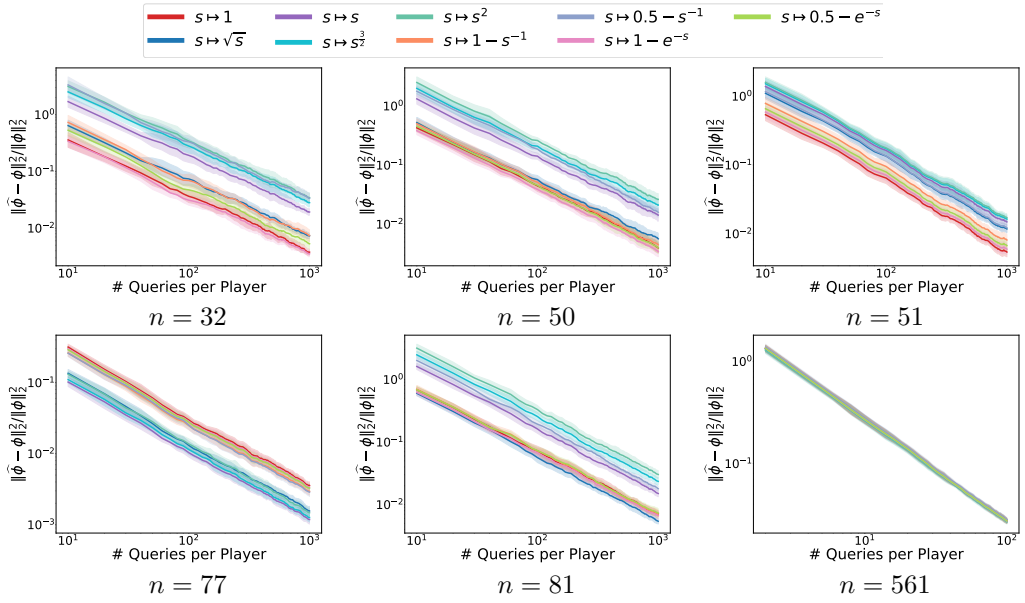


Figure 5: All curves are averaged over 10 random seeds, with standard deviations reported. Each curve corresponds to a particular choice of f, with the estimate produced according to Eq. (9).

B More Experimental Details and Results

Table 1 summarizes all datasets used in our experiments and reports the specified depth of the gradient-boosting decision trees used for training. More experimental results can be found in Figures 5, 6, and 7.

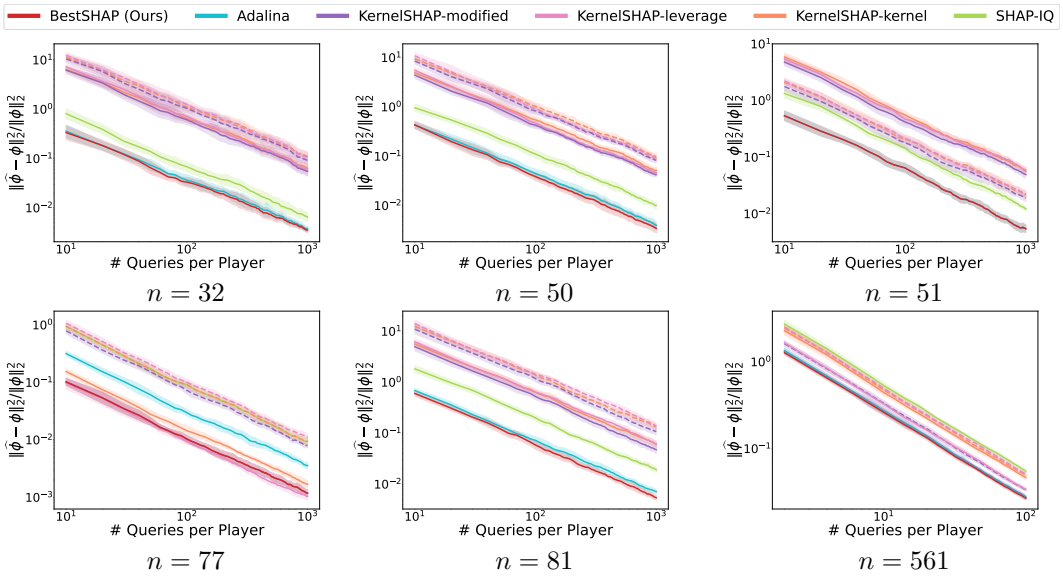


Figure 6: All curves are averaged over 10 random seeds, with standard deviations reported. The dashed corresponds to the use of $\alpha = 0$ in Eq. (7), whereas the solid line uses $\alpha = \frac{u_{[n]} - u_\emptyset}{n}$.

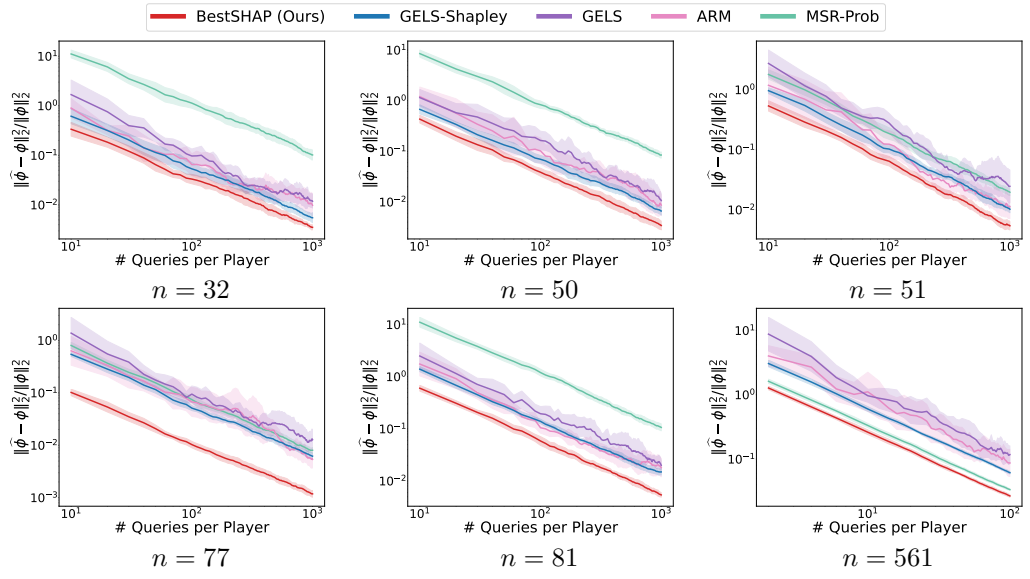


Figure 7: All curves are averaged over 10 random seeds, with standard deviations reported.