PREDICTABLE SIZE-CONDITIONAL LINEAR CONTROL VARIATES FOR UNBIASED SEMI-VALUE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Shapley values and more generally, semi-values attribute a model's prediction to individual players, such as features or training points. Utility assigns a score to every possible subset of players; however, exact computation is exponential, so practice relies on unbiased sampling with variance reduction. ADALINA recently removed $\log n$ factor from the query complexity, at linear time and space per utility query. However, its single global scalar control variate cannot capture variation among same-size coalitions, and fitting it on the samples it corrects induces an $O(1/T)$ bias. Observing that this within-size variation is often close to linear, we propose SLICE, which adds a size-wise mean and a size-wise linear control variate with closed-form centering, both fit only on past batches and held fixed for the current one. SLICE is unbiased at finite sample size, keeps ADALINA's linear cost and $\log n$-free query complexity, and, with population-optimal coefficients, does not increase MSE over its global-only counterpart. Across 12 datasets and 18 semi-value settings, SLICE matches or outperforms every baseline, with a median error reduction of $2.1\times$.

1 INTRODUCTION

The Shapley value (Shapley et al., 1953) is the unique allocation of a cooperative game's total value satisfying linearity, symmetry, the null-player, and efficiency, while semi-values (Dubey et al., 1981) drop efficiency. Given n players and a utility function $U(S)$ on coalitions $S \subseteq [n]$, player i's allocation is a weighted average of marginal contributions $U(S \cup \{i\}) - U(S)$ across coalitions. In machine learning, this is a foundational tool for feature attribution and data valuation (Lundberg & Lee, 2017). Exact computation requires evaluating $U(S)$ over an exponential number of coalitions, so practice relies on unbiased Monte Carlo approximation.

Prior work has progressed along two axes (Section F). The first reduces estimator variance: stratified sampling with optimized allocation across strata (Castro et al., 2017), and regression that fits a surrogate to $U(S)$ (Lundberg & Lee, 2017), improved by paired sampling (Covert & Lee, 2021). The second improves estimator's query complexity. Early analyses carry an $O(\log n)$ factor, arising from a coordinate-wise union bound (Li & Yu, 2024b) or matrix concentration (Musco & Witter, 2025). Li et al. (2026a) remove this factor with a dimension-free vector concentration inequality, and propose ADALINA , which adds a single scalar control variate at $\Theta(n)$ space and $O(n)$ time per sample.

ADALINA's global scalar has two gaps. First, fitting it on the same samples it corrects biases the estimator (Li et al. (2026a); see also Proposition 4.1). Second, a single scalar cannot capture within-size heterogeneity, i.e., variation among coalitions of the same size. Regression-adjusted estimators (Witter et al., 2025) capture richer structure by fitting $V \approx U$, but they require $\phi(V)$ to be efficiently computable, which restricts V to tractable families such as linear or tree models; the linear case requires solving an $n \times n$ least-squares system. In our diagnostic experiment across ten tabular datasets, reallocating the sampling distribution across coalition sizes yields little benefit beyond ADALINA's optimized distribution $q^\star$, while within-size variation is close to linear in coalition membership.

Motivated by this, we propose SLICE (Size-conditional LInear Control-variate Estimator), which augments the global correction with a size-wise mean term and a size-wise linear term, while preserving finite-sample unbiasedness, dimension-free concentration, and linear time and space complexity. All coefficients are estimated from past batches and held fixed during the current one,

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.

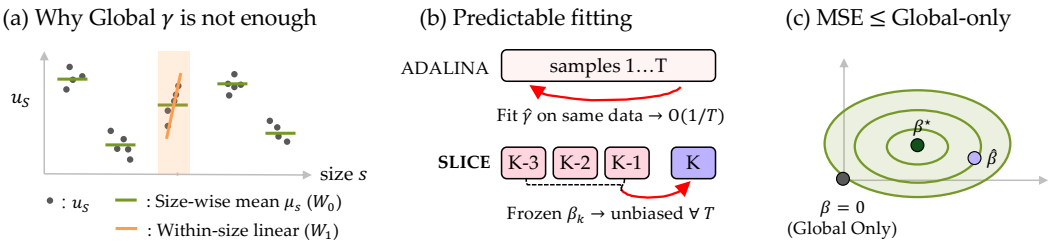

Figure 1: **Overview of SLICE.** (a) A global γ misses the size-wise mean μ_s and within-size linear variation of u_S; SLICE removes both via W_0, W_1. (b) Fitting $\hat{\gamma}$ on the same T samples causes $O(1/T)$ bias; SLICE freezes β_k fit on past batches, staying unbiased. (c) Any $\hat{\beta}$ in the shaded region (minimizer $\beta^\star$) has MSE no larger than global-only ($\beta = 0$).

thus eliminates the dependence between a coefficient and the observation it multiplies, removing ADALINA's $O(1/T)$ bias. Such predictable coefficients, however, make per-sample contributions dependent across batches, so the independent and identically distributed (i.i.d.) vector concentration argument of Li et al. (2026a) no longer applies. We instead clip all coefficients to a bounded range, and apply a dimension-free Hilbert-space martingale inequality (Pinelis, 1994).

Our contributions are as follows:

- Our diagnostic study motivates a size-conditional control variate, whose linear term admits closed-form centering, keeping $\mathcal{O}(n)$ time and space per query (Lemma 4.2).
- We prove that SLICE is unbiased at every fixed sample size T for any predictable coefficients, including those produced by warm-up, clipping, and online gating (Theorem 5.1).
- Since predictable coefficients break the i.i.d. assumption used by ADALINA, we recover its $\log n$-free $\mathcal{O}(n\epsilon^{-2}\log(1/\delta))$ query complexity with a Hilbert-space martingale inequality (Theorem 5.4). With population-optimal coefficients, the added control variates do not increase MSE over the predictable global-only baseline (Theorem 5.6), and we characterize when online coefficients retain this (Theorem 5.8).
- Across 12 datasets and 18 semi-value settings, SLICE matches or outperforms the best baseline, with a median error reduction of $2.1\times$, and our ablation shows that two controls are complementary.

2 PRELIMINARIES

2.1 SEMI-VALUES

Let $N = [n] = \{1, \dots, n\}$, and let $U : 2^N \to \mathbb{R}$ be a utility function satisfying $\|U\|_\infty := \max_{S\subseteq N} |U(S)| \le C$. We write $u_S = U(S)$ and $\mathbf{1}_S \in \{0,1\}^n$ denote the membership vector of S.

Definition 2.1 (Semi-value). *Let $p_1, \dots, p_n \ge 0$ satisfy the normalization conditions of a semi-value (Dubey et al., 1981), and define $m_s := \binom{n-1}{s-1} p_s$ for $s = 1, \dots, n$. The semi-value of U is the vector $\phi(U) \in \mathbb{R}^n$ whose i-th coordinate is $\phi_i(U) = \sum_{S\subseteq N\setminus\{i\}} p_{|S|+1}\big(U(S\cup\{i\}) - U(S)\big)$.*

Remark 2.2 (Instances of semi-values). *The Shapley value (Shapley et al., 1953) sets $p_s = (s-1)!(n-s)!/n!$, equivalently $m_s = 1/n$ for every s. The weighted Banzhaf value (Wang & Jia, 2023) is indexed by $w \in (0,1)$ via $p_s = w^{s-1}(1-w)^{n-s}$, recovering the (unweighted) Banzhaf value at $w = 1/2$. Beta Shapley (Kwon & Zou, 2022) is indexed by $\alpha, \beta \ge 1$ via $p_s = B(s+\beta-1, n-s+\alpha)/B(\alpha,\beta)$, where B denotes the Beta function; $(\alpha,\beta) = (1,1)$ recovers the Shapley value.*

For each $\varnothing \ne S \subsetneq [n]$, define $w_S \in \mathbb{R}^n$ by $(w_S)_i = p_{|S|}\mathbf{1}\{i \in S\} - p_{|S|+1}\mathbf{1}\{i \notin S\}$. Then

$$\phi(U) = \underbrace{\textstyle\sum_{\emptyset\subsetneq S\subsetneq N} u_S w_S}_{\phi^\circ(U)} + \underbrace{(m_n u_N - m_1 u_\emptyset)\mathbf{1}_n}_{b_\partial(U)}, \tag{1}$$

$b_\partial(U)$ is exactly computable using two utility queries. We therefore focus on estimating $\phi^\circ(U)$.

2.2 Optimal sampling distribution

Li et al. (2026a) estimate $\phi^\circ(U)$ by first sampling a cardinality s from a distribution q supported on the active set $\mathcal{I} := \{s \in \{1, \dots, n-1\} : m_s^2/s + m_{s+1}^2/(n-s) > 0\}$, and then sampling S uniformly among subsets of cardinality s. For $s = |S|$, define the importance vector

$$(z_S(q))_i = \frac{n}{q_s}\left(\frac{m_s}{s}\mathbf{1}\{i \in S\} - \frac{m_{s+1}}{n-s}\mathbf{1}\{i \notin S\}\right) \tag{2}$$

that satisfies $\mathbb{E}_q[u_S z_S(q)] = \phi^\circ(U)$. Conditionally on the sampled cardinality,

$$\mathbb{E}[z_S(q) \mid |S| = s] = e_s(q)\mathbf{1}_n, \qquad e_s(q) := \frac{m_s - m_{s+1}}{q_s}, \tag{3}$$

and averaging over s gives

$$\mathbb{E}[z_S(q)] = \sum_{s \in \mathcal{I}} q_s e_s(q)\mathbf{1}_n = c_0\mathbf{1}_n, \qquad c_0 := m_1 - m_n. \tag{4}$$

Therefore, $z_S - c_0\mathbf{1}_n$ is globally zero-mean, whereas $z_S - e_s\mathbf{1}_n$ is zero-mean conditional on s. Symmetric semi-values have $m_1 = m_n$, so $c_0 = 0$. The sampling distribution governs both the estimator's variance and its query complexity through the single scalar

$$D(q) := \sum_{s \in \mathcal{I}} \frac{n}{q_s}\left(\frac{m_s^2}{s} + \frac{m_{s+1}^2}{n-s}\right). \tag{5}$$

By the Cauchy–Schwarz inequality, $D(q) \geq D^\star$ for every q, with equality if and only if

$$q_s = q_s^\star \propto \sqrt{n\left(m_s^2/s + m_{s+1}^2/(n-s)\right)}, \qquad D^\star := \left(\sum_{s \in \mathcal{I}} \sqrt{n\big(m_s^2/s + m_{s+1}^2/(n-s)\big)}\right)^2. \tag{6}$$

We now fix $q = q^\star$ and denote $z_S := z_S(q^\star)$. This setting ensures that the norm of the importance vector is uniform for all $\emptyset \subsetneq S \subsetneq N$, satisfying $\|z_S\|_2 = Z := \sqrt{nD^\star}$. In Section 5, we present a concentration bound using $D^\star$ and Z (Equation (20)).

2.3 Filtration and predictability

In our setting, we observe coalitions $S_1, S_2, \dots$ sequentially. Let $\mathcal{F}_{t-1}$ denote the σ-field generated by everything observed strictly before S_t is sampled. A quantity is *predictable* if it is $\mathcal{F}_{t-1}$-measurable (Meyer, 1962), ensuring that a zero-mean control variate remains zero-mean when multiplied by an adaptive coefficient, a property we exploit throughout Sections 4 and 5.

3 Motivations

Structural Decomposition of the Utility. Set functions are often approximated by low-order terms (Owen, 1972; Grabisch, 1997; Fumagalli et al., 2023). Following this view, we write

$$U(S) = \mu_{|S|} + \langle b^\star, \mathbf{1}_S \rangle + \varepsilon_S, \tag{7}$$

where $\mu_s := \mathbb{E}_{q^\star}[u_S \mid |S| = s]$ is the size-wise mean, $b^\star := \arg\min_{b \perp \mathbf{1}_n} \mathbb{E}_{q^\star}\left[(u_S - \mu_{|S|} - \langle b, \mathbf{1}_S \rangle)^2\right]$ is a single linear effect shared across sizes, and ε_S is the remaining interaction. Since adding $c\mathbf{1}_n$ to $b^\star$ only shifts $\mu_{|S|}$, we remove this freedom by requiring $b^\star \perp \mathbf{1}_n$.

Variance Reduction Mechanisms. Stratified sampling reallocates samples across coalition sizes (Castro et al., 2017; Maleki et al., 2013), but it needs pilot samples and cannot reduce variance within a size. Regression-adjusted Monte Carlo (Witter et al., 2025) fits a surrogate $V \approx U$ and samples only the residual, but $\phi(V)$ must be computable in closed form, and the surrogate must be trained. We instead use control variates (Asmussen & Glynn, 2007): zero-mean terms under $q^\star$ with closed-form centering. They keep $q^\star$, need no surrogate training, and admit unbiasedness and an exponential tail bound (Section 5).

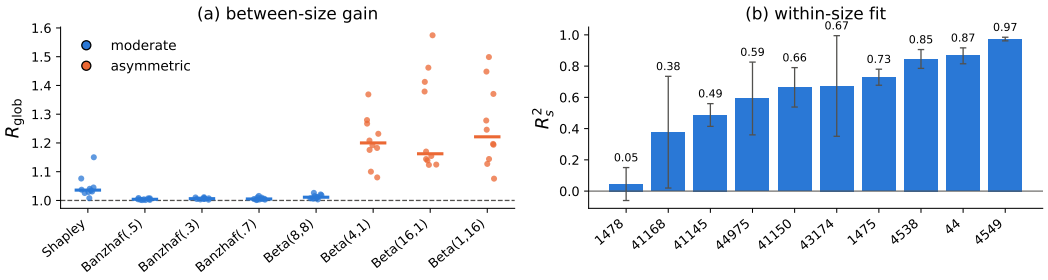

Figure 2: Diagnostic experiment on ten tabular datasets. (a) Neyman-optimal between-size reallocation gain R_{glob} by semi-value family; each point is one dataset, the bar marks the median. (b) Within-size linear R_s^2 by dataset (x-axis: OpenML dataset ID, see Table 1); error bars show the standard deviation over sizes.

Prior Control Variates. ADALINA (Li et al., 2026a) subtracts a single constant γ. One scalar cannot match μ_s for every size, nor remove $\langle b^\star, \mathbf{1}_S \rangle$, which varies within a size. Fitting $\widehat{\gamma}_T$ on the same samples also causes an $\mathcal{O}(1/T)$ bias (Proposition 4.1). Model-based control variates (Goldwasser & Hooker, 2024) and surrogate-based estimators (Witter et al., 2025) can capture richer structure, but require an approximating model. SLICE adds a size-wise mean and a size-wise linear term, both with closed-form centering and $\mathcal{O}(n)$ cost per query.

Empirical Variance Diagnostics. We use two diagnostics to fix the design of SLICE: whether to modify the sampling distribution, and which functional form to use for the within-size control. Both are computed on gradient-boosted tree games from ten tabular datasets (Section G.1) under Shapley, weighted Banzhaf, and Beta-Shapley values. **(1) Sampling distribution:** Let $v_s := \mathbb{E}[(u_S - \gamma^\star)^2 \mid |S| = s] = \mathrm{Var}(u_S \mid |S| = s) + (\mu_s - \gamma^\star)^2$ be the squared deviation that the global mean $\gamma^\star = \mathbb{E}_{q^\star}[u_S]$ leaves at size s. Since $\|z_S(q^\star)\|_2$ is constant, the largest possible reduction of the second moment from reallocating samples across sizes, ignoring the cost of estimating v_s, is the factor $R_{\text{glob}} := \mathbb{E}_{q^\star}[v]/(\mathbb{E}_{q^\star}[\sqrt{v}])^2 \geq 1$ (Cauchy–Schwarz). In Figure 2(a), $R_{\text{glob}} \leq 1.15$ for the Shapley value and moderately asymmetric semi-values, and its median reaches only 1.16–1.20 under extreme Beta-Shapley asymmetry. We therefore keep $q^\star$, which also keeps $\|z_S\|_2 = Z$ constant as required by Theorem 5.4. Since reallocation cannot remove the drift $(\mu_s - \gamma^\star)^2$ within each sample, we remove it with W_0. **(2) Functional form:** For each size s among the sizes with sufficiently large $\binom{n}{s}$, we fit a ridge regression of u_S on $\mathbf{1}_S$ and report its 5-fold out-of-fold R_s^2. High R_s^2 values indicate linear structure, while low values are inconclusive. In Figure 2(b), the mean R_s^2 has a median of about 0.66 across datasets (up to 0.97), so a first-order model explains much of the within-size variation in most games. We therefore use a linear control, which, unlike a high-capacity surrogate, has a closed-form conditional mean computable in $\mathcal{O}(n)$ time (Lemma 4.2). Since this diagnostic fits each size separately while W_1 shares one b across sizes, we measure the actual gains of W_0 and W_1 in Section G.3. These gains vary across settings, and because the cost of estimating coefficients remains even when the gain is zero (Section 5.4), we gate each coefficient online.

4 METHOD

4.1 OVERVIEW

Figure 1 is an overview of SLICE. Recall from Section 2.2 that $z_S(q)$ satisfies $\mathbb{E}_q[u_S z_S(q)] = \phi^\circ(U)$. Li et al. (2026a) reduce the variance of $u_S z_S$ by using $u_S z_S - \gamma \cdot z_S$. However, a single scalar γ can remove variation in one direction, whereas the residual variance in u_S comes from two sources: variation *across* coalition sizes $s = |S|$ (different sizes have different average utilities), and variation *within* a fixed size (different players contribute differently even at the same s). SLICE addresses both by adding two zero-mean correction terms:

$$\underbrace{u_S z_S - \gamma \cdot z_S}_{\text{global mean (ADALINA)}} - \underbrace{\beta_0 \cdot W_0}_{\text{cardinality-wise mean}} - \underbrace{\beta_1 \cdot W_1}_{\text{cardinality-wise linear}} . \tag{8}$$

For asymmetric semi-values, a correction handles $c_0 = m_1 - m_n \neq 0$ in Equation (9) below.

Fitting γ, W_0, W_1, and their coefficients β_0, β_1 from the same samples they correct makes each coefficient depend on the observations it multiplies. For example, using $\widehat{\gamma}_T = \frac{1}{T}\sum_{t=1}^{T} u_{S_t}$ in $\widehat{\phi}_T^{\mathrm{Ada},\circ} = \frac{1}{T}\sum_{t=1}^{T}(u_{S_t} - \widehat{\gamma}_T)z_{S_t}$ on the same T samples gives the following bias:

Proposition 4.1 (Finite-sample bias (Li et al., 2026a, proof of Theorem 4.1))**.** *For a symmetric semi-value and i.i.d. samples from $q^\star$, $\mathbb{E}[\widehat{\phi}_T^{\mathrm{Ada},\circ}] = (1 - \frac{1}{T})\phi^\circ(U)$, so $\mathrm{Bias}(\widehat{\phi}_T^{\mathrm{Ada},\circ}) = -\frac{1}{T}\phi^\circ(U)$, which vanishes as $T \to \infty$ and is generally nonzero at any finite T.*

The problem is not the control variate but fitting and evaluating it on the same observations. SLICE therefore fits every coefficient on earlier batches and freezes it for the current one (Section 4.5; Figure 1b), which yields unbiasedness at every finite T (Theorem 5.1).

4.2 PREDICTABLE SIZE-CONDITIONAL CONTROL VARIATES

Let $\mathcal{F}_{t-1}$ be the filtration of Section 2.3, containing all information observed strictly before S_t is sampled. Conditionally on $\mathcal{F}_{t-1}$ we sample $s_t \sim q^\star$ and then S_t uniformly among subsets of cardinality s_t. Throughout, a real number γ_t, a per-cardinality mean estimate $\bar{u}_{s,t}$ (a predictable estimate of $\mathbb{E}[u_S \mid |S| = s]$), a shared linear-effect vector $b_t \in \mathbb{R}^n$ (a predictable estimate of the additive component $b^\star$ of U in Equation (7)), and coefficients $\beta_{0,s,t}, \beta_{1,s,t} \in \mathbb{R}$ are all assumed $\mathcal{F}_{t-1}$-measurable for every $s \in \mathcal{I}$. In the implementation they are recomputed once at the start of each batch from previous-batch statistics and then held fixed for every sample in that batch (Section 4.5).

Global term. Let $s_t = |S_t|$. Define

$$R_t := u_{S_t} z_{S_t} - \gamma_t(z_{S_t} - c_0 \mathbf{1}_n). \tag{9}$$

By Equation (4), $\mathbb{E}[\gamma_t(z_{S_t} - c_0\mathbf{1}_n) \mid \mathcal{F}_{t-1}] = 0$, so R_t is ADALINA's global correction with a γ_t.

Cardinality-wise mean term. This control variate isolates the part of the variance explained by the drift between the global mean γ_t and the cardinality-s mean $\bar{u}_{s,t}$:

$$W_{0,t} := (\bar{u}_{s_t,t} - \gamma_t)\big(z_{S_t} - e_{s_t}(q^\star)\mathbf{1}_n\big). \tag{10}$$

By Equation (3), $\mathbb{E}[W_{0,t} \mid \mathcal{F}_{t-1}, s_t] = 0$.

Cardinality-wise linear term. This control variate isolates the part of the within-cardinality variance explained linearly by coalition membership. Writing $\ell_t(S) := \langle b_t, \mathbf{1}_S \rangle$, define

$$W_{1,t} := \ell_t(S_t) z_{S_t} - c_{s_t}^{\mathrm{lin}}(b_t), \qquad c_s^{\mathrm{lin}}(b_t) := \mathbb{E}[\ell_t(S) z_S \mid |S| = s], \tag{11}$$

so that $\mathbb{E}[W_{1,t} \mid \mathcal{F}_{t-1}, s_t] = 0$.

4.3 CLOSED-FORM LINEAR CENTERING

Computing $c_s^{\mathrm{lin}}(b_t)$ by definition would require Monte Carlo approximation. Below shows it is instead available in closed form. For a fixed size s, write $\pi_1 = s/n$ and $\pi_2 = s(s-1)/\big(n(n-1)\big)$. z_S in Section 2.2 can be reformulated as $(z_S)_i = c_- + (c_+ - c_-)\mathbf{1}\{i \in S\}$, where $c_+ = \frac{n}{q_s^\star}\frac{m_s}{s}$ and $c_- = -\frac{n}{q_s^\star}\frac{m_{s+1}}{n-s}$. Let $d_s := c_+ - c_-$ and $B_t := \mathbf{1}_n^\top b_t$.

Lemma 4.2 (Closed-form linear centering)**.** *For every $s \in \mathcal{I}$,*

$$c_s^{\mathrm{lin}}(b_t) = K_{s,t}\mathbf{1}_n + d_s(\pi_1 - \pi_2)b_t, \qquad K_{s,t} = c_-\pi_1 B_t + d_s\pi_2 B_t. \tag{12}$$

Hence $W_{1,t}$ is computable in $O(n)$ time and $O(n)$ memory, without Monte Carlo approximation of its conditional expectation.

4.4 THE SLICE ESTIMATOR

Let $\beta_{s,t} := (\beta_{0,s,t}, \beta_{1,s,t})^\top \in \mathcal{B} := [0, \bar{\beta}]^2$ for a fixed clipping level $\bar{\beta} > 0$. Combining terms of Equation (8), the round-t contribution is

$$G_t := R_t - \beta_{0,s_t,t} W_{0,t} - \beta_{1,s_t,t} W_{1,t}, \tag{13}$$

Algorithm 1 SLICE (**S**ize-conditional **LI**near **C**ontrol-variate **E**stimator)

Require: Semi-value weights $m_1, \dots, m_n$; utility U; number T of interior queries; batch size B; ridge parameter λ; clipping level $\bar{\beta}$

Evaluate $u_\emptyset$ and u_N, Compute $q^\star$, $c_0 = m_1 - m_n$, and $e_s(q^\star)$
Initialize global and per-stratum sufficient statistics
for each batch k **do**
  Using only previous batches, compute $\gamma_k, \bar{u}_{s,k}, b_k, \beta_{0,s,k}, \beta_{1,s,k}$ (Section 4.5)
  for each sample in batch k **do**
    Sample $s \sim q^\star$, then S uniformly with $|S| = s$
    Query $u_S = U(S)$ and compute z_S
    Compute R, W_0, W_1 via Equations (9) to (12)
    Accumulate $G = R - \beta_{0,s,k} W_0 - \beta_{1,s,k} W_1$
    Update sufficient statistics for use in the next batch
  end for
end for
return $T^{-1} \sum_{t=1}^{T} G_t + (m_n u_N - m_1 u_\emptyset)\mathbf{1}_n$

and, for a number T of interior samples, we add the boundary term $b_\partial(U)$ of Equation (1) to obtain

$$\widehat{\phi}_T^{\text{SLICE}} := \frac{1}{T} \sum_{t=1}^{T} G_t + b_\partial(U). \tag{14}$$

4.5 ONLINE COEFFICIENT FITTING

Index batches by $k = 1, 2, \dots$, and let τ_k denote the time index of the first sample in batch k, so that $M_{k-1} := \tau_k - 1$ is the number of interior samples completed before batch k begins ($M_0 = 0$). For every sample t in batch k we set $\gamma_t = \gamma_k$, $\bar{u}_{s,t} = \bar{u}_{s,k}$, $b_t = b_k$, and $\beta_{j,s,t} = \beta_{j,s,k}$. We now specify how each of these is computed from batches $1, \dots, k-1$ only.

Global mean γ_k.

$$\gamma_k = \frac{u_\emptyset + u_N + \sum_{r<\tau_k} u_{S_r}}{M_{k-1} + 2}. \tag{15}$$

Since $|u_S| \le C$ (Section 2.1), γ_k is an average of terms bounded by C and thus $|\gamma_k| \le C$.

Cardinality mean $\bar{u}_{s,k}$. Let $N_{s,k}$ be the number of samples with $s_r = s$ among the first M_{k-1} interior samples. Then $\bar{u}_{s,k} = \big(\sum_{r<\tau_k,\, s_r=s} u_{S_r}\big)/N_{s,k}$ if $N_{s,k} \ge 1$, and $\bar{u}_{s,k} = \gamma_k$ otherwise.

Linear coefficient b_k. With $x_S := \mathbf{1}_S - \frac{s}{n}\mathbf{1}_n$ (so $\mathbb{E}[x_S \mid |S| = s] = 0$), the shared linear-effect vector is the diagonal (moment-based) regression estimate of $u_{S_r} - \gamma_k$ on x_{S_r} over all interior samples observed before k:

$$b_k = \frac{\sum_{r<\tau_k} u_{S_r}\, x_{S_r} - \gamma_k \sum_{r<\tau_k} x_{S_r}}{\sum_{r<\tau_k} \frac{s_r}{n}\big(1 - \frac{s_r}{n}\big) + \lambda}. \tag{16}$$

Unlike γ_k, the ridge estimate in Equation (16) is not automatically bounded, so we rescale it, if necessary, to satisfy $\|b_k\|_1 \le 2C_k$, where $C_k := \max\big(|u_\emptyset|, |u_N|, \max_{r<\tau_k} |u_{S_r}|\big)$.

Stratum statistics for β. For every active s, let $A_{00,s,k}, A_{11,s,k}, A_{01,s,k}$ and $r_{0,s,k}, r_{1,s,k}$ denote the cumulative inner products $\sum \|W_0\|_2^2$, $\sum \|W_1\|_2^2$, $\sum \langle W_0, W_1\rangle$, $\sum \langle R, W_0\rangle$, $\sum \langle R, W_1\rangle$ over all samples $r < \tau_k$ with $s_r = s$ (using the R, W_0, W_1 values computed when each such sample was processed), and let $\Sigma_{s,k} := \sum_{r<\tau_k,\, s_r=s} \|R_r\|_2^2$.

Coefficients $\beta_{0,s,k}, \beta_{1,s,k}$**.** With a small ridge term $\varepsilon_{s,k} > 0$ (added only to avoid division by zero when $A_{00,s,k} + A_{11,s,k}$ is small), the joint least-squares fit solves

$$\begin{pmatrix} \widetilde{\beta}_{0,s,k} \\ \widetilde{\beta}_{1,s,k} \end{pmatrix} = \begin{pmatrix} A_{00,s,k} + \varepsilon_{s,k} & A_{01,s,k} \\ A_{01,s,k} & A_{11,s,k} + \varepsilon_{s,k} \end{pmatrix}^{-1} \begin{pmatrix} r_{0,s,k} \\ r_{1,s,k} \end{pmatrix}, \tag{17}$$

i.e., with $\Delta_{s,k} := (A_{00,s,k} + \varepsilon_{s,k})(A_{11,s,k} + \varepsilon_{s,k}) - A_{01,s,k}^2$, We activate $\widetilde{\beta}_{j,s,k}$ only once M_{k-1} and $N_{s,k}$ exceed a threshold and it clears a growing threshold inspired by upper-confidence-bound exploration rules (Auer et al., 2002): with $t_{j,s,k} := r_{j,s,k}\sqrt{N_{s,k}/(A_{jj,s,k}\,\Sigma_{s,k})}$ and $u_{s,k} := \sqrt{\frac{2(1+N_{s,k})}{N_{s,k}} \log \frac{\sqrt{1+N_{s,k}}}{\delta_{\text{gate}}}}$,

$$\beta_{j,s,k} = \begin{cases} \text{clip}(\widetilde{\beta}_{j,s,k}, 0, \bar{\beta}), & t_{j,s,k} > u_{s,k}, \\ 0, & \text{otherwise}, \end{cases} \tag{18}$$

where $\text{clip}(x, 0, \bar{\beta}) := \min(\max(x, 0), \bar{\beta})$ guarantees $\beta_{s,k} \in \mathcal{B}$, as required by Section 5.

Remark 4.3 (Interpretation of the gate)**.** *Equation (18) implements a data-dependent gating rule for detecting signal, not a formally verified anytime-valid test, left for future work. Nonetheless, Theorem 5.1 requires that $\beta_{j,s,t}$ be predictable, regardless of the gate's validity.*

5 Theory

5.1 Finite-sample unbiasedness

Theorem 5.1 (Finite-sample unbiasedness)**.** *For every t, suppose that $\gamma_t, \bar{u}_{s,t}, b_t, \beta_{0,s,t}, \beta_{1,s,t}$ are $\mathcal{F}_{t-1}$-measurable. Then, for every semi-value, every utility function, and every deterministic $T \geq 1$,*

$$\mathbb{E}[\widehat{\phi}_T^{\text{SLICE}}] = \phi(U).$$

This conclusion remains valid under warm-up, clipping, ridge regularization, and data-dependent gating, provided their outputs are predictable.

Remark 5.2 (Deterministic horizons)**.** *Theorem 5.1 concerns a deterministic number of samples. A time-uniform concentration inequality may remain valid at a stopping time, but fixed-horizon unbiasedness does not imply unbiasedness under arbitrary data-dependent stopping.*

5.2 Dimension-free concentration

We impose the boundedness conditions used by the implementation:

$$|u_S| \leq C, \qquad |\gamma_t| \leq C, \qquad |\bar{u}_{s,t}| \leq C, \qquad \|b_t\|_1 \leq 2C, \qquad 0 \leq \beta_{j,s,t} \leq \bar{\beta}. \tag{19}$$

Define

$$H := (3 + 8\bar{\beta})C\sqrt{nD^\star}. \tag{20}$$

Proposition 5.3 (Bounded contributions)**.** *Under Equation (19), $\|G_t\|_2 \leq H$. Moreover, with $X_t := G_t - \phi^\circ(U)$, we have $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$ and $\|X_t\|_2 \leq 2H$.*

Since (X_t) is a bounded martingale-difference sequence in $\mathbb{R}^n$, a dimension-free Hilbert-space Hoeffding inequality (Pinelis, 1994, Thm. 3.5) (restated in Section C.1 in Appendix) gives:

Theorem 5.4 (Dimension-free concentration)**.** *Under the assumptions of Theorem 5.1 and proposition 5.3, there is a universal constant $c > 0$ such that*

$$\mathbb{P}\left(\left\|\widehat{\phi}_T^{\text{SLICE}} - \phi(U)\right\|_2 \geq \epsilon\right) \leq 2\exp\left(-c\frac{T\epsilon^2}{(3+8\bar{\beta})^2 C^2 nD^\star}\right). \tag{21}$$

The bound contains no multiplicative factor depending on the ambient dimension outside $nD^\star$.

Corollary 5.5 (Query complexity)**.** *For a universal constant $C_0 > 0$, it is sufficient to take*

$$T \geq C_0 \frac{(3+8\bar{\beta})^2 C^2 nD^\star}{\epsilon^2} \log \frac{2}{\delta}$$

to guarantee $\mathbb{P}\left(\left\|\widehat{\phi}_T^{\text{SLICE}} - \phi(U)\right\|_2 \geq \epsilon\right) \leq \delta$. Thus, if $D^\star = O(1)$ and $\bar{\beta} = O(1)$, then $T = O\left(\frac{nC^2}{\epsilon^2} \log \frac{1}{\delta}\right)$. Including the exact boundary evaluations, the total number of utility queries is $T + 2$.

5.3 Provable Variance Reduction from the Cardinality-Wise Controls

Let $\widehat{\phi}_T^{(0)}$ denote the *global-only* estimator obtained by fixing $\beta_{0,s,t} \equiv \beta_{1,s,t} \equiv 0$ from Equation (13):

$$\widehat{\phi}_T^{(0)} := \frac{1}{T}\sum_{t=1}^{T} R_t + b_\partial(U), \tag{22}$$

Unlike ADALINA's $\widehat{\gamma}_T$, its predictable γ_t makes it unbiased at every finite T (Theorem 5.1). We now show that MSE can be further reduced by W_0 and W_1 with appropriately chosen coefficients.

For a fixed time t and stratum s, let $\mathcal{W}_t := (W_{0,t}, W_{1,t})$ denote the $\mathbb{R}^n$-valued control pair. Define

$$A_{s,t} := \mathbb{E}\left[\begin{pmatrix}\langle W_{0,t}, W_{0,t}\rangle & \langle W_{0,t}, W_{1,t}\rangle \\ \langle W_{1,t}, W_{0,t}\rangle & \langle W_{1,t}, W_{1,t}\rangle\end{pmatrix} \middle| \mathcal{F}_{t-1}, s_t = s\right], \tag{23}$$

$$h_{s,t} := \mathbb{E}\left[\begin{pmatrix}\langle R_t, W_{0,t}\rangle \\ \langle R_t, W_{1,t}\rangle\end{pmatrix} \middle| \mathcal{F}_{t-1}, s_t = s\right]. \tag{24}$$

Because the controls have conditional mean zero, R_t can be replaced by $R_t - \phi^\circ(U)$ in Equation (24). $A^\dagger$ denotes the Moore–Penrose pseudoinverse, and the population-optimal coefficient denotes $\beta^\star_{s,t}$:

$$\beta^\star_{s,t} := A^\dagger_{s,t} h_{s,t} \tag{25}$$

Theorem 5.6 (Non-negativity of the W_0, W_1 variance reduction)**.** *For every s, t, $h_{s,t} \in \mathrm{Range}(A_{s,t})$. Let $\widehat{\phi}^\star_T$ denote the estimator that replaces $\beta_{0,s_t,t}, \beta_{1,s_t,t}$ in Equation (13) by $\beta^\star_{s_t,t} = A^\dagger_{s_t,t} h_{s_t,t}$. Then*

$$\mathrm{MSE}(\widehat{\phi}^\star_T) - \mathrm{MSE}(\widehat{\phi}^{(0)}_T) = -\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{s\in\mathcal{I}} q^\star_s h_{s,t}^\top A^\dagger_{s,t} h_{s,t}\right] \le 0. \tag{26}$$

Thus, adding W_0 and W_1 with the population-optimal coefficients does not increase the MSE of $\widehat{\phi}^{(0)}_T$.

Corollary 5.7 (Strict variance reduction)**.** *The inequality in Equation (26) is strict if $\mathbb{P}(h_{s,t}^\top A^\dagger_{s,t} h_{s,t} > 0) > 0$ for at least one pair (s,t) with $q^\star_s > 0$.*

5.4 Online Coefficients

For arbitrary predictable coefficients $\beta_{s,t}$, let $\|v\|^2_{A_{s,t}} := v^\top A_{s,t} v$.

Theorem 5.8 (MSE decomposition)**.** *Relative to the population-optimal coefficients of Equation (25),*

$$\begin{aligned}\mathrm{MSE}(\widehat{\phi}^{\text{SLICE}}_T) - \mathrm{MSE}(\widehat{\phi}^{(0)}_T) = &-\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{s\in\mathcal{I}} q^\star_s h_{s,t}^\top A^\dagger_{s,t} h_{s,t}\right] \\ &+ \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{s\in\mathcal{I}} q^\star_s \left\|\beta_{s,t} - \beta^\star_{s,t}\right\|^2_{A_{s,t}}\right].\end{aligned} \tag{27}$$

The first term is the always nonpositive variance reduction attainable with the population-optimal coefficients, and the second is a nonnegative coefficient-estimation penalty.

Corollary 5.9 (Finite-sample online dominance condition)**.** *Online* SLICE *(using the estimated $\beta_{s,t}$ of Section 4.5) attains an MSE no larger than $\widehat{\phi}^{(0)}_T$ whenever*

$$\sum_{t=1}^{T}\mathbb{E}\left[\sum_{s} q^\star_s \left\|\beta_{s,t} - \beta^\star_{s,t}\right\|^2_{A_{s,t}}\right] \le \sum_{t=1}^{T}\mathbb{E}\left[\sum_{s} q^\star_s h_{s,t}^\top A^\dagger_{s,t} h_{s,t}\right].$$

5.5 Computational Complexity

Proposition 5.10 (Time and space complexity)**.** *Each utility query requires $O(n)$ arithmetic to construct z_S, W_0, W_1, and update the sufficient statistics. A coefficient refit uses $O(n)$ arithmetic plus $O(1)$ work per stratum. The total complexity for T samples is therefore $O(Tn)$, and the total memory requirement is $O(n)$.*

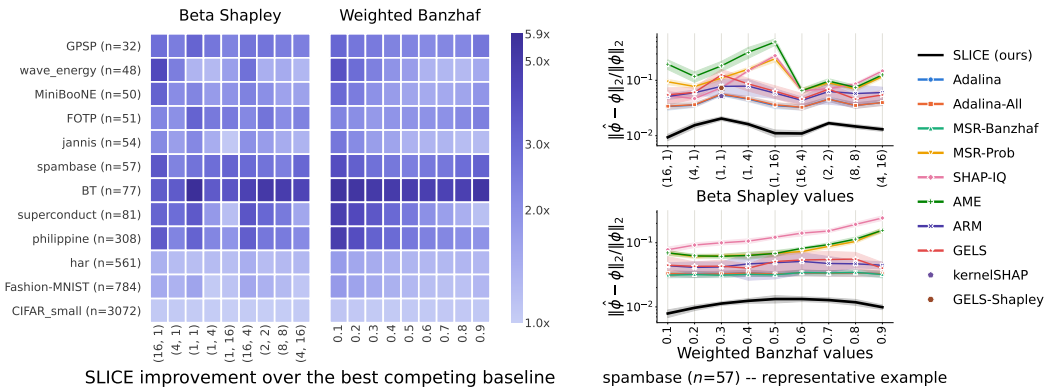

Figure 3: SLICE matches or outperforms the best baseline across 12 datasets and 18 semi value settings, with a representative example (right) showing the gap in detail.

6 EXPERIMENTS

Setup. Following Li et al. (2026a), we use 12 OpenML benchmark datasets (Vanschoren et al., 2014) with n features as players, and the same gradient-boosted tree-ensemble utilities, for which exact semi-values are computable in closed form (Li et al., 2026b) and serve as ground truth. Each method receives $T = nQ$ interior queries, where Q is the number of queries per player (Table 2), and results are averaged over 10 seeds. SLICE uses a single hyperparameter setting across all datasets and semi-values. Details on datasets, utilities, and hyperparameters are in Sections G.1 and G.2.

Baselines. We compare SLICE against baselines for semi-value estimation: SHAP-IQ (Fumagalli et al., 2023), Adalina and Adalina-All (Li et al., 2026a), GELS (Li & Yu, 2024a), ARM (Kolpaczki et al., 2024), and AME (Lin et al., 2022), for which we substitute the linear-space variant of Li & Yu (2024b). We also include MSR-Prob, the sampling distribution of (Witter et al., 2025) as adapted into a semi-value baseline by (Li et al., 2026a). Three further baselines are compared only against the specific semi-values they target: MSR-Banzhaf (Wang & Jia, 2023) approximates weighted Banzhaf values, while unbiased KernelSHAP (Chen et al., 2026) and GELS-Shapley (Li & Yu, 2024a) are specialized for the Shapley value. (For detailed related work, please refer to Section F.)

Results. We report the relative ℓ_2 error $\|\hat{\phi} - \phi\|_2 / \|\phi\|_2$, averaged over 10 seeds. As shown in Figure 3, SLICE matches or outperforms the best competing baseline in all 216 settings (12 datasets $\times$ 18 semi-values), with a median error reduction of $2.1\times$. The margin is largest on BT at the Shapley value ($5.9\times$) and smallest on CIFAR_small at $(\alpha, \beta) = (1, 16)$, where the two are nearly tied. Full per-dataset curves are in Figures 4 and 5.

To further validate our design, we conduct an ablation that disables W_0 or W_1 on three datasets with R_s^2 ranging from 0.05 to 0.85 (Table 3 in Section G.3). The predictable global-only baseline performs on par with ADALINA in most settings. W_1 usually contributes more for Beta-Shapley$(16, 1)$, and the two are complementary. For example, on superconduct in Shapley value, W_0 and W_1 alone reduce error from 0.085 to 0.041 and 0.077, while combining them reaches 0.024.

7 CONCLUSION

We propose SLICE, a predictable size-conditional linear control-variate estimator that augments ADALINA's global correction with a size-wise mean term and a within-size linear term, both fit only from past batches. Across twelve datasets and three semi-value families, SLICE matches or outperforms estimation error among all baselines tested. Theoretically, we prove that SLICE is unbiased at every deterministic finite horizon (Theorem 5.1), preserves ADALINA's dimension-free query complexity (Theorem 5.4), and, with population-optimal coefficients, does not increase MSE over the global-only baseline (Theorem 5.6). For future work, we plan to replace the current UCB gate with an anytime-valid test for within-size linear signal. (Details are in Section H)

AI USE STATEMENT

In this work, we used Claude Opus 5.5 as a writing and editing assistant to translate and refine the text. All technical content was reviewed and verified by the authors, who take full responsibility for its correctness.

REFERENCES

Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Sebastian Bordt and Ulrike von Luxburg. From shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, pp. 709–745. PMLR, 2023.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.

Javier Castro, Daniel Gómez, Elisenda Molina, and Juan Tejada. Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research*, 82:180–188, 2017.

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. *Advances in Neural Information Processing Systems*, 38:164943–164995, 2026.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *International conference on artificial intelligence and statistics*, pp. 3457–3465. PMLR, 2021.

Pradeep Dubey, Abraham Neyman, and Robert James Weber. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. Shap-iq: Unified approximation of any-order shapley interactions. *Advances in Neural Information Processing Systems*, 36:11515–11551, 2023.

Jeremy Goldwasser and Giles Hooker. Stabilizing estimates of shapley values with control variates. In *World Conference on Explainable Artificial Intelligence*, pp. 416–439. Springer, 2024.

Michel Grabisch. K-order additive discrete fuzzy measures and their representation. *Fuzzy sets and systems*, 92(2):167–189, 1997.

Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 38, pp. 13246–13255, 2024.

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 8780–8802, 2022. URL https://proceedings.mlr.press/v151/kwon22a.html.

Weida Li and Yaoliang Yu. Robust data valuation with weighted banzhaf values. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=u359tNBpxF.

Weida Li and Yaoliang Yu. Faster approximation of probabilistic and distributional values via least squares. In *International Conference on Learning Representations*, volume 2024, pp. 51182–51216, 2024a.

Weida Li and Yaoliang Yu. One sample fits all: Approximating all probabilistic values simultaneously and efficiently. *Advances in Neural Information Processing Systems*, 37:58309–58340, 2024b.

Weida Li, Yaoliang Yu, and Bryan Kian Hsiang Low. Adalina: Adaptive linear approximation for the shapley value and beyond. In *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.

Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. Measuring the effect of training data on deep learning predictions via randomized experiments. In *International Conference on Machine Learning*, pp. 13468–13504. PMLR, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

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(1):56–67, 2020.

Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*, 2013.

Paul-André Meyer. A decomposition theorem for supermartingales. *Illinois Journal of Mathematics*, 6(2):193–205, 1962.

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.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Lloyd S Shapley et al. A value for n-person games. 1953.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International conference on artificial intelligence and statistics*, pp. 6388–6421. PMLR, 2023.

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.

A Auxiliary Identities and Centering Derivations

A.1 Conditional mean of the importance vector

Fix $s \in \mathcal{I}$. For each player i,

$$\mathbb{P}(i \in S \mid |S| = s) = \frac{s}{n}, \qquad \mathbb{P}(i \notin S \mid |S| = s) = \frac{n-s}{n}.$$

Using Equation (2),

$$\begin{aligned}\mathbb{E}[(z_S)_i \mid |S| = s] &= \frac{n}{q_s}\left[\frac{m_s}{s}\frac{s}{n} - \frac{m_{s+1}}{n-s}\frac{n-s}{n}\right] \\ &= \frac{m_s - m_{s+1}}{q_s} = e_s(q).\end{aligned}$$

Therefore, $\mathbb{E}[z_S \mid |S| = s] = e_s(q)\mathbf{1}_n$.

Averaging over s gives

$$\begin{aligned}\mathbb{E}[z_S] &= \sum_{s\in\mathcal{I}} q_s e_s(q)\mathbf{1}_n \\ &= \sum_{s=1}^{n-1}(m_s - m_{s+1})\mathbf{1}_n \\ &= (m_1 - m_n)\mathbf{1}_n = c_0\mathbf{1}_n.\end{aligned}$$

The second equality remains valid after omitting inactive strata because an inactive stratum has $m_s = m_{s+1} = 0$.

A.2 Proof of Lemma 4.2

Fix s and suppress the time index. Let

$$I_i := \mathbf{1}\{i \in S\}, \qquad L := \langle b, \mathbf{1}_S\rangle = \sum_{j=1}^{n} b_j I_j.$$

Under uniform sampling among subsets of size s,

$$\mathbb{E}[I_i] = \pi_1 = \frac{s}{n}, \qquad \mathbb{E}[I_i I_j] = \pi_2 = \frac{s(s-1)}{n(n-1)} \quad \text{for } i \neq j.$$

Let $B = \sum_j b_j$. Then $\mathbb{E}[L] = \sum_j b_j \mathbb{E}[I_j] = \pi_1 B$. Moreover,

$$\begin{aligned}\mathbb{E}[LI_i] &= b_i\mathbb{E}[I_i] + \sum_{j\neq i} b_j \mathbb{E}[I_i I_j] \\ &= b_i\pi_1 + (B - b_i)\pi_2 \\ &= \pi_2 B + (\pi_1 - \pi_2)b_i.\end{aligned}$$

Because $(z_S)_i = c_- + d_s I_i$, we obtain

$$\begin{aligned}\mathbb{E}[L(z_S)_i \mid |S| = s] &= c_-\mathbb{E}[L] + d_s\mathbb{E}[LI_i] \\ &= c_-\pi_1 B + d_s\pi_2 B + d_s(\pi_1 - \pi_2)b_i.\end{aligned}$$

Defining $K_s = c_-\pi_1 B + d_s\pi_2 B$ gives $c_s^{\mathrm{lin}}(b) = K_s\mathbf{1}_n + d_s(\pi_1 - \pi_2)b$, which is Equation (12).

B Bias and Unbiasedness

B.1 Proof of Proposition 4.1

For a symmetric semi-value,

$$\mathbb{E}[z_S] = 0.$$

Let

$$Y_t := u_{S_t} z_{S_t}, \qquad \widehat{\gamma}_T = \frac{1}{T}\sum_{k=1}^{T} u_{S_k}.$$

Then

$$\widehat{\phi}_T^{\mathrm{Ada},\circ} = \frac{1}{T}\sum_{t=1}^{T} Y_t - \widehat{\gamma}_T \left(\frac{1}{T}\sum_{t=1}^{T} z_{S_t}\right).$$

The first term has expectation $\phi^\circ(U)$. For the second term,

$$\mathbb{E}\left[\widehat{\gamma}_T \left(\frac{1}{T}\sum_{t=1}^{T} z_{S_t}\right)\right] = \frac{1}{T^2}\sum_{k=1}^{T}\sum_{t=1}^{T} \mathbb{E}[u_{S_k} z_{S_t}].$$

Independence gives $\mathbb{E}[u_{S_k} z_{S_t}] = \mathbb{E}[u_{S_k}]\mathbb{E}[z_{S_t}] = 0$ for $k \neq t$, while $\mathbb{E}[u_{S_t} z_{S_t}] = \phi^\circ(U)$ for $k = t$. There are T diagonal terms, so $\mathbb{E}[\widehat{\gamma}_T(\frac{1}{T}\sum_{t=1}^{T} z_{S_t})] = \frac{1}{T}\phi^\circ(U)$. Therefore, $\mathbb{E}[\widehat{\phi}_T^{\mathrm{Ada},\circ}] = (1 - \frac{1}{T})\phi^\circ(U)$, which proves the claim.

B.2 Proof of Theorem 5.1

By the importance-sampling identity, $\mathbb{E}[u_{S_t} z_{S_t} \mid \mathcal{F}_{t-1}] = \phi^\circ(U)$, because the sampling distribution at round t is $q^\star$ and is independent of the past.

Since γ_t is $\mathcal{F}_{t-1}$-measurable,

$$\begin{aligned}\mathbb{E}[\gamma_t(z_{S_t} - c_0\mathbf{1}_n) \mid \mathcal{F}_{t-1}] &= \gamma_t\left(\mathbb{E}[z_{S_t}] - c_0\mathbf{1}_n\right)\\ &= 0.\end{aligned}$$

Therefore, $\mathbb{E}[R_t \mid \mathcal{F}_{t-1}] = \phi^\circ(U)$.

Next, conditional on $\mathcal{F}_{t-1}$ and $s_t = s$,

$$\begin{aligned}\mathbb{E}[W_{0,t} \mid \mathcal{F}_{t-1}, s_t = s] &= (\bar{u}_{s,t} - \gamma_t)\left(\mathbb{E}[z_{S_t} \mid s_t = s] - e_s\mathbf{1}_n\right)\\ &= 0.\end{aligned}$$

Likewise, by the definition of $c_s^{\mathrm{lin}}(b_t)$,

$$\begin{aligned}\mathbb{E}[W_{1,t} \mid \mathcal{F}_{t-1}, s_t = s] &= \mathbb{E}[\ell_t(S_t) z_{S_t} \mid \mathcal{F}_{t-1}, s_t = s] - c_s^{\mathrm{lin}}(b_t)\\ &= 0.\end{aligned}$$

Averaging over s_t preserves these zero conditional expectations, using the tower property $\mathbb{E}[\beta_{j,s_t,t} W_{j,t} \mid \mathcal{F}_{t-1}] = \mathbb{E}\big[\mathbb{E}[\beta_{j,s_t,t} W_{j,t} \mid \mathcal{F}_{t-1}, s_t] \mid \mathcal{F}_{t-1}\big]$ together with the fact that, conditionally on $\mathcal{F}_{t-1}$ and $s_t = s$, $\beta_{j,s,t}$ is a fixed ($\mathcal{F}_{t-1}$-measurable) number that factors out of the inner conditional expectation, which is zero.

The coefficients $\{\beta_{0,s,t}\}_{s\in\mathcal{I}}$ and $\{\beta_{1,s,t}\}_{s\in\mathcal{I}}$ are $\mathcal{F}_{t-1}$-measurable. Consequently, $\mathbb{E}[G_t \mid \mathcal{F}_{t-1}] = \phi^\circ(U)$. Taking expectations and averaging over $t = 1, \ldots, T$, $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} G_t] = \phi^\circ(U)$. Finally, adding the deterministic boundary term gives

$$\mathbb{E}[\widehat{\phi}_T^{\mathrm{SLICE}}] = \phi^\circ(U) + b_\partial(U) = \phi(U).$$

B.3 Batchwise predictability

Let batch k begin after all observations in batches $1, \dots, k-1$ have been incorporated. Let $\mathcal{G}_{k-1}$ be the sigma-field generated by those observations. All coefficient refits at the beginning of batch k are deterministic functions of the maintained sufficient statistics and hence are $\mathcal{G}_{k-1}$-measurable.

For any sample t in batch k, $\mathcal{G}_{k-1} \subseteq \mathcal{F}_{t-1}$. Therefore, the batch-start coefficients are also $\mathcal{F}_{t-1}$-measurable. Holding them fixed within the batch satisfies the predictability condition of Theorem 5.1.

C Concentration and Query Complexity

C.1 Hilbert-valued martingale Hoeffding inequality (Pinelis, 1994, Thm. 3.5)

Let $X_1, \dots, X_T$ be martingale differences in a separable Hilbert space and suppose $\|X_t\| \le L_t$ almost surely, where L_t is fixed. Then there is a universal constant $c_{\mathrm{H}} > 0$, independent of the Hilbert-space dimension, such that

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T} X_t\right\| \ge x\right) \le 2\exp\left(-c_{\mathrm{H}}\frac{x^2}{\sum_{t=1}^{T} L_t^2}\right).$$

C.2 Proof of Proposition 5.3

Recall $\|z_S\|_2 = Z$ from Equation (6). By Equation (4) and Jensen's inequality, $\|c_0\mathbf{1}_n\|_2 = \|\mathbb{E}[z_S]\|_2 \le \mathbb{E}\,\|z_S\|_2 = Z$, so $\|z_S - c_0\mathbf{1}_n\|_2 \le 2Z$. Using $|u_S| \le C$ and $|\gamma_t| \le C$,

$$\begin{aligned}\|R_t\|_2 &\le |u_{S_t}|\,\|z_{S_t}\|_2 + |\gamma_t|\,\|z_{S_t} - c_0\mathbf{1}_n\|_2 \\ &\le CZ + 2CZ = 3CZ.\end{aligned}$$

Conditional Jensen gives $\|e_s\mathbf{1}_n\|_2 = \|\mathbb{E}[z_S \mid |S| = s]\|_2 \le Z$. Since $|\bar{u}_{s,t} - \gamma_t| \le 2C$,

$$\begin{aligned}\|W_{0,t}\|_2 &\le 2C\left(\|z_{S_t}\|_2 + \|e_{s_t}\mathbf{1}_n\|_2\right) \\ &\le 4CZ.\end{aligned}$$

For the linear control, $|\ell_t(S)| = |\,\langle b_t, \mathbf{1}_S\rangle\,| \le \|b_t\|_1 \le 2C$, so $\|\ell_t(S)z_{S_t}\|_2 \le 2CZ$. Again using conditional Jensen,

$$\begin{aligned}\left\|c_s^{\mathrm{lin}}(b_t)\right\|_2 &= \|\mathbb{E}[\ell_t(S)z_S \mid |S| = s]\|_2 \\ &\le \mathbb{E}[|\ell_t(S)|\,\|z_S\|_2 \mid |S| = s] \\ &\le 2CZ.\end{aligned}$$

Hence, $\|W_{1,t}\|_2 \le 4CZ$.

Because the two coefficients are bounded by $\bar{\beta}$,

$$\begin{aligned}\|G_t\|_2 &\le \|R_t\|_2 + \bar{\beta}\,\|W_{0,t}\|_2 + \bar{\beta}\,\|W_{1,t}\|_2 \\ &\le (3 + 8\bar{\beta})CZ = H.\end{aligned}$$

By Theorem 5.1, $\phi^\circ(U) = \mathbb{E}[G_t \mid \mathcal{F}_{t-1}]$, so conditional Jensen gives $\|\phi^\circ(U)\|_2 \le H$. Consequently, $\|X_t\|_2 = \|G_t - \phi^\circ(U)\|_2 \le 2H$.

C.3 Proof of Theorem 5.4

The boundary term cancels from the estimation error:

$$\widehat{\phi}_T^{\,\mathrm{SLICE}} - \phi(U) = \frac{1}{T}\sum_{t=1}^{T}(G_t - \phi^\circ(U)) = \frac{1}{T}\sum_{t=1}^{T} X_t.$$

By Proposition 5.3, $(X_t, \mathcal{F}_t)$ is an $\mathbb{R}^n$-valued martingale-difference sequence satisfying $\|X_t\|_2 \le 2H$. Applying Section C.1 with $L_t = 2H$ and $x = T\epsilon$ gives

$$\begin{aligned}\mathbb{P}\left(\left\|\widehat{\phi}_T^{\text{SLICE}} - \phi(U)\right\|_2 \ge \epsilon\right) &= \mathbb{P}\left(\left\|\sum_{t=1}^{T} X_t\right\|_2 \ge T\epsilon\right) \\ &\le 2\exp\left(-c_{\text{H}}\frac{T^2\epsilon^2}{4TH^2}\right) \\ &= 2\exp\left(-\frac{c_{\text{H}}}{4}\frac{T\epsilon^2}{H^2}\right).\end{aligned}$$

Substituting

$$H^2 = (3 + 8\bar{\beta})^2 C^2 n D^\star$$

proves Equation (21), with $c = c_{\text{H}}/4$.

Solving the right-hand side for T proves Corollary 5.5.

D Variance Reduction from the Cardinality-Wise Controls

D.1 Martingale orthogonality

Let $X_t^{(0)} := R_t - \phi^\circ(U)$. Then $\mathbb{E}[X_t^{(0)} \mid \mathcal{F}_{t-1}] = 0$. For $r < t$, the vector $X_r^{(0)}$ is $\mathcal{F}_{t-1}$-measurable, and therefore

$$\begin{aligned}\mathbb{E}\left\langle X_r^{(0)}, X_t^{(0)}\right\rangle &= \mathbb{E}\left[\left\langle X_r^{(0)}, \mathbb{E}[X_t^{(0)} \mid \mathcal{F}_{t-1}]\right\rangle\right] \\ &= 0.\end{aligned}$$

The same argument applies to every predictable controlled increment. Thus, MSEs of the sample average decompose as sums of one-step conditional second moments.

D.2 The range condition

Fix s, t and suppress the indices. Let A and h be defined by Equations (23) and (24). The matrix A is positive semidefinite because, for every $v \in \mathbb{R}^2$, $v^\top A v = \mathbb{E}\left[\|v_0 W_0 + v_1 W_1\|_2^2 \,\middle|\, \mathcal{F}_{t-1}, s_t = s\right] \ge 0$.

Suppose $v \in \text{Null}(A)$. Then $0 = v^\top A v = \mathbb{E}\left[\|v_0 W_0 + v_1 W_1\|_2^2 \,\middle|\, \mathcal{F}_{t-1}, s_t = s\right]$. A non-negative random variable with zero conditional expectation is zero almost surely, so $v_0 W_0 + v_1 W_1 = 0$ almost surely conditionally on $(\mathcal{F}_{t-1}, s_t = s)$. It follows that $v^\top h = \mathbb{E}\left[\langle R_t, v_0 W_0 + v_1 W_1\rangle \,|\, \mathcal{F}_{t-1}, s_t = s\right] = 0$. Hence $h \perp \text{Null}(A)$. For a symmetric finite-dimensional matrix, $\text{Range}(A) = \text{Null}(A)^\perp$, so $h \in \text{Range}(A)$.

D.3 Proof of Theorem 5.6

Fix s, t and let $X_t(\beta) := R_t - \phi^\circ(U) - \beta_0 W_{0,t} - \beta_1 W_{1,t}$. Expanding the conditional second moment gives

$$\begin{aligned}\mathbb{E}[\|X_t(\beta)\|_2^2 \mid \mathcal{F}_{t-1}, s_t = s] = \mathbb{E}[\|R_t - \phi^\circ(U)\|_2^2 \mid \mathcal{F}_{t-1}, s_t = s] \\ + \beta^\top A_{s,t}\beta - 2h_{s,t}^\top \beta.\end{aligned} \tag{28}$$

Because $h_{s,t} \in \text{Range}(A_{s,t})$, every unconstrained minimizer has the form $A_{s,t}^\dagger h_{s,t} + v$ with $v \in \text{Null}(A_{s,t})$. We choose the minimum-norm representative $\beta_{s,t}^\star = A_{s,t}^\dagger h_{s,t}$. Substitution into the quadratic term gives $(\beta_{s,t}^\star)^\top A_{s,t}\beta_{s,t}^\star - 2h_{s,t}^\top \beta_{s,t}^\star = -h_{s,t}^\top A_{s,t}^\dagger h_{s,t}$.

After averaging over s_t,

$$\mathbb{E}[\left\|X_t(\beta^\star_{s_t,t})\right\|_2^2] - \mathbb{E}[\left\|X_t^{(0)}\right\|_2^2] = -\mathbb{E}\left[\sum_{s\in\mathcal{I}} q_s^\star h_{s,t}^\top A_{s,t}^\dagger h_{s,t}\right].$$

Martingale orthogonality eliminates cross-time terms. Dividing the sum of the one-step differences by T^2 yields Equation (26).

Since $A_{s,t}^\dagger$ is positive semidefinite on $\mathrm{Range}(A_{s,t})$, $h_{s,t}^\top A_{s,t}^\dagger h_{s,t} \geq 0$. This proves the weak inequality of Theorem 5.6. Strict improvement (Corollary 5.7) follows if this quantity is positive with positive probability for at least one active (s,t).

E Online Coefficients

E.1 Proof of Theorem 5.8

For fixed s,t, abbreviate $A = A_{s,t}$, $h = h_{s,t}$, and $\beta^\star = A^\dagger h$. Because $h \in \mathrm{Range}(A)$, $A\beta^\star = h$. For an arbitrary β,

$$\begin{aligned}\beta^\top A\beta - 2h^\top\beta &= \beta^\top A\beta - 2(\beta^\star)^\top A\beta \\ &= (\beta-\beta^\star)^\top A(\beta-\beta^\star) - (\beta^\star)^\top A\beta^\star \\ &= \|\beta-\beta^\star\|_A^2 - h^\top A^\dagger h.\end{aligned}$$

Substituting this identity into Equation (28), averaging over s_t, summing over time, and using martingale orthogonality gives Equation (27).

The finite-sample condition in Corollary 5.9 follows by requiring the nonnegative penalty term not to exceed the variance-reduction term.

F Extended Related Work

F.1 Sampling-Based Approximation of Semi-Values

Because exact semi-value computation requires exponential utility queries, existing literature relies on sampling-based approximations of $\phi(U)$. Castro et al. (2009) introduce uniform random-permutation sampling for Shapley values, which Castro et al. (2017) refine via Neyman-optimal size-stratification. Maleki et al. (2013) propose size-stratified coalition sampling whose allocation minimizes a Hoeffding-type error bound. Chen et al. (2026) subsequently unify KernelSHAP(Lundberg & Lee, 2017), unbiased KernelSHAP (Covert & Lee, 2021), and LeverageSHAP (Musco & Witter, 2025) as instances of a single sketched-regression or approximate matrix-vector-multiplication estimator, providing the first non-asymptotic sample-complexity guarantees for KernelSHAP itself. Instead of sampling marginal contributions, SVARM and Stratified SVARM (Kolpaczki et al., 2024) estimate Shapley values from sampled coalition values, with each query updating the estimates for many players. Wang & Jia (2023) formalize this maximum-sample-reuse (MSR) principle for the Banzhaf value, which Li & Yu (2023) extended to weighted Banzhaf values.

Complementary research targets the dimension dependence in sample complexity. Conventional estimators establish ℓ_2-error guarantees through coordinate-wise concentration, introducing an extra $\log n$ factor via union bounds. Li & Yu (2024b) propose OFA, which reuses each sampled subset for all players and remove the additional polylogarithmic overheads, although the union-bound $\log n$ factor remains. Li et al. (2026a) eliminate this factor by applying a dimension-free vector concentration inequality to a linear-space unbiased counterpart of OFA, and propose ADALINA, which further reduces the MSE via an adaptively estimated scalar control variate. Our proposed framework, SLICE, builds upon ADALINA's $q^\star$-sampling distribution and Hilbert-space martingale inequality (Pinelis, 1994). Crucially, we incorporate size-conditional linear controls (Section 4) that, with population-optimal coefficients, do not increase mean-squared error (Theorem 5.6), and do not reintroduce a $\log n$ factor (Theorem 5.4).

F.2 Regression-Based and Surrogate Approximations

An alternative paradigm substitutes the evaluation of the utility function with an analytic surrogate. When U is constrained to tree ensembles, Lundberg et al. (2020) compute expected marginal contributions in polynomial time by recursively traversing trees conditioned on node statistics, a principle generalized to arbitrary semi-values by Li et al. (2026b). However, this approach requires access to internal model topologies and does not apply to black-box utilities.

For general utilities, Lin et al. (2022) formulate semi-value estimation as linear regression (AME) by fitting linear surrogates to sampled coalitions. We adopt the linear-space formulation of Li & Yu (2024b) as a baseline to circumvent the requirement of storing the complete design matrix. Li & Yu (2024a) generalize this least-squares perspective to broader probabilistic and distributional values (GELS), including an efficiency-based Shapley value variant (GELS-Shapley), that calculates Shapley values via additive-efficient normalization instead of introducing a null player.

Witter et al. (2025) extend this surrogate framework to arbitrary high-capacity models (e.g., gradient-boosted decision trees, such as TreeMSR variant) trained offline using k-fold cross-fitting under a fixed sampling budget. However, debiasing non-linear surrogates discards the underlying martingale structure. Our linear control W_1 (Lemma 4.2) admits closed-form conditional expectations given coalition cardinality, preserving true zero-mean martingale difference sequences (Theorem 5.1) and exponential tail bounds in $1/\delta$ (Section C.1). Consequently, SLICE targets the order-0 and order-1 structure of U, trading the flexibility of learned surrogates for closed-form centering, unbiasedness at every fixed T, and exponential concentration.

F.3 Interaction Indices and Functional Decompositions of Set Functions

Equation (7) corresponds to an order-0 and order-1 truncation of the classical hierarchical decomposition of cooperative games. Owen (1972) introduces the multilinear extension, expressing the Shapley value as a line integral of utility gradients over the continuous unit cube. Sundararajan et al. (2020) generalize this formulation to order-k Taylor expansions via the Shapley-Taylor interaction index. Tsai et al. (2023) define FAITH-SHAP via a weighted least-squares fit to the value function, unique under the standard Shapley axioms, which Bordt & von Luxburg (2023) later reinterpret as a truncated polynomial approximation—paralleling the Taylor-truncation view of Shapley-Taylor. Fumagalli et al. (2023) unify these formulations into a single sampling estimator for arbitrary-order interactions (SHAP-IQ).

In this hierarchical decomposition, Equation (7) corresponds to an order-0 and order-1 truncation, where the residual ε_S absorbs all interaction terms of order ≥ 2. Our diagnostic experiments (Section 3) indicate that an order-1 model explains a substantial fraction of within-size variance in most of the games we study (median $R_s^2 \approx 0.66$), although not in all (e.g., $R_s^2 = 0.05$ on har). This motivates the order-1 truncation as a low-cost choice rather than establishing its optimality.

G Extended Experiments

G.1 Datasets and Utility Functions

Models. Following Li et al. (2026a), we train a `GradientBoostingClassifier` or `GradientBoostingRegressor` (scikit-learn) with 10 trees and the dataset-specific maximum depth listed in Table 1, on an 80/20 train/test split (stratified for classification) with base random seed 2026. For the two regression targets with large output scale, we standardize y with a scaler fit on the training split.

Utility functions. Players correspond to input features. For each dataset, we fix one target test instance x, sampled uniformly at random from the held-out split (seed offset by dataset id); for classification tasks, we additionally sample one target class uniformly at random and attribute that class's logit. Given the trained ensemble f, the utility of a coalition $S \subseteq [n]$ is the path-dependent conditional expectation

$$U(S) = \mathbb{E}_{X_{[n]\setminus S}}\big[f(x_S, X_{[n]\setminus S})\big],$$

Table 1: Summary of the 12 OpenML datasets, following the benchmark suite of Li et al. (2026a).

Dataset	ID	#Instances	#Features	Domain	Task	Depth
FOTP	1475	6,118	51	tabular	classification (6)	20
GPSP	4538	9,873	32	tabular	classification (5)	10
MiniBooNE	41150	130,064	50	tabular	classification (2)	20
philippine	41145	5,832	308	tabular	classification (2)	15
spambase	44	4,601	57	tabular	classification (2)	15
jannis	41168	83,733	54	tabular	classification (4)	40
har	1478	10,299	561	tabular	classification (6)	20
superconduct	43174	21,263	81	tabular	regression	10
wave_energy	44975	72,000	48	tabular	regression	30
BT	4549	583,250	77	tabular	regression	20
Fashion-MNIST	40996	70,000	784	image	classification (10)	30
CIFAR-10-small	40926	20,000	3,072	image	classification (10)	20

Table 2: Interior-sample budget (queries per player, Q) by dataset, following Li et al. (2026a). The total number of interior queries is $T = nQ$.

Datasets	Q
FOTP, GPSP, MiniBooNE, philippine, spambase, jannis, superconduct, wave_energy, BT	1000
har, Fashion-MNIST	100
CIFAR-10-small	20

evaluated recursively from each tree's own branch statistics (Lundberg et al., 2020) rather than by Monte Carlo sampling over the data distribution.

Ground truth. We use tree-ensemble utilities because their semi-values admit exact polynomial-time computation, which provides ground truth for measuring estimation error. Specifically, $\phi(U)$ is computed via the recursive tree-traversal algorithm of Li et al. (2026b), which generalizes path-dependent TreeSHAP (Lundberg et al., 2020) to the full semi-value family through Gauss–Legendre quadrature over the mixing distribution.

G.2 Sample Budget and Hyperparameters

All reported results (Figures 3 to 5) are averaged over 10 independent random seeds per (dataset, semi-value) configuration.

Unless otherwise noted, SLICE uses the following fixed hyperparameters, held constant across all datasets and semi-value settings: batch size $B = 100$ (the coefficient-refit cadence of Algorithm 1), ridge parameter $\lambda = 1$, clipping level $\bar{\beta} = 1$, and the UCB gate of Equation (18) with $\delta_{\text{gate}} = 0.1$. The per-stratum coefficients $\beta_{0,s,k}, \beta_{1,s,k}$ are held at 0 until $M_{k-1} \geq 8(n-1)$ interior samples and $N_{s,k} \geq 8$ per-stratum samples have been observed.

G.3 Ablation Study on Control Variate

To isolate the individual contribution of W_0 and W_1, we compare five estimators on three datasets spanning the range of within-size linear structure R_s^2 measured in Section 3: $R_s^2 = 0.05$ for har, 0.67 for superconduct, and 0.85 for GPSP, so within-size linear structure increases from har to GPSP. The five estimators are ADALINA; the predictable global-only baseline $\widehat{\phi}_T^{(0)}$ (Equation (22)); $\widehat{\phi}_T^{(0)}$ with only W_0 active ($\beta_{1,s,t} \equiv 0$ in Equation (13)); $\widehat{\phi}_T^{(0)}$ with only W_1 active ($\beta_{0,s,t} \equiv 0$ in Equation (13)); and SLICE (Equation (13) with both coefficients active). That is, the four estimators other than ADALINA are all Algorithm 1, differing only in which coefficient is held at 0. We report the relative ℓ_2 error at the final checkpoint, averaged over 10 seeds, for the Shapley value and the asymmetric Beta-Shapley$(16, 1)$ value. The budget at each dataset's final checkpoint follows Table 2.

Table 3: Ablation study on the effect of W_0 and W_1 under Shapley (1,1) and Beta-Shapley (16,1) settings. SLICE (ours) combines global, W_0 and W_1.

dataset	**Shapley (1,1)**					**Beta-Shapley (16,1)**				
	Adalina	global	$+W_0$	$+W_1$	**SLICE**	Adalina	global	$+W_0$	$+W_1$	**SLICE**
har	0.1652	0.1653	0.1571	0.1473	**0.1378**	0.1309	0.1311	0.1288	0.0969	**0.0939**
superconduct	0.0844	0.0845	0.0413	0.0772	**0.0241**	0.0324	0.0327	0.0312	0.0139	**0.0103**
GPSP	0.0613	0.0613	0.0347	0.0540	**0.0195**	0.0329	0.0340	0.0339	0.0119	**0.0117**

As shown in Table 3, SLICE achieves the lowest error. While both W_0 and W_1 individually reduce error, their combination yields a lower error than using either one alone. Which component contributes more significantly depends on the dataset and the semi-value. Specifically, W_0 dominates in the Shapley value calculation for the superconduct dataset, whereas W_1 plays the dominant role in Beta-Shapley$(16, 1)$.

H Limitation and Future Work

Theorem 5.6 holds for the population-optimal coefficients $\beta^{\star}_{s,t}$ of Equation (25). For the coefficients estimated online in Section 4.5, the MSE of SLICE does not exceed that of $\widehat{\phi}_T^{(0)}$ only when the condition of Corollary 5.9 holds. Furthermore, the gate in Equation (18) is a heuristic threshold based on upper-confidence-bound rules and is not an anytime-valid hypothesis test. The unbiasedness in Theorem 5.1 holds for a deterministic number of samples T and does not extend to data-dependent stopping times.

For future work, we plan to derive sufficient conditions on the batch size and the warm-up length under which the condition of Corollary 5.9 holds for every finite T. We also plan to replace the gate with an anytime-valid test for within-size linear signal and to combine it with a time-uniform concentration inequality, which would allow data-dependent stopping. Finally, we plan to extend W_1 to pairwise interaction terms, whose conditional expectation given $|S| = s$ also admits a closed form, and to evaluate SLICE on black-box utilities such as retraining-based data valuation.

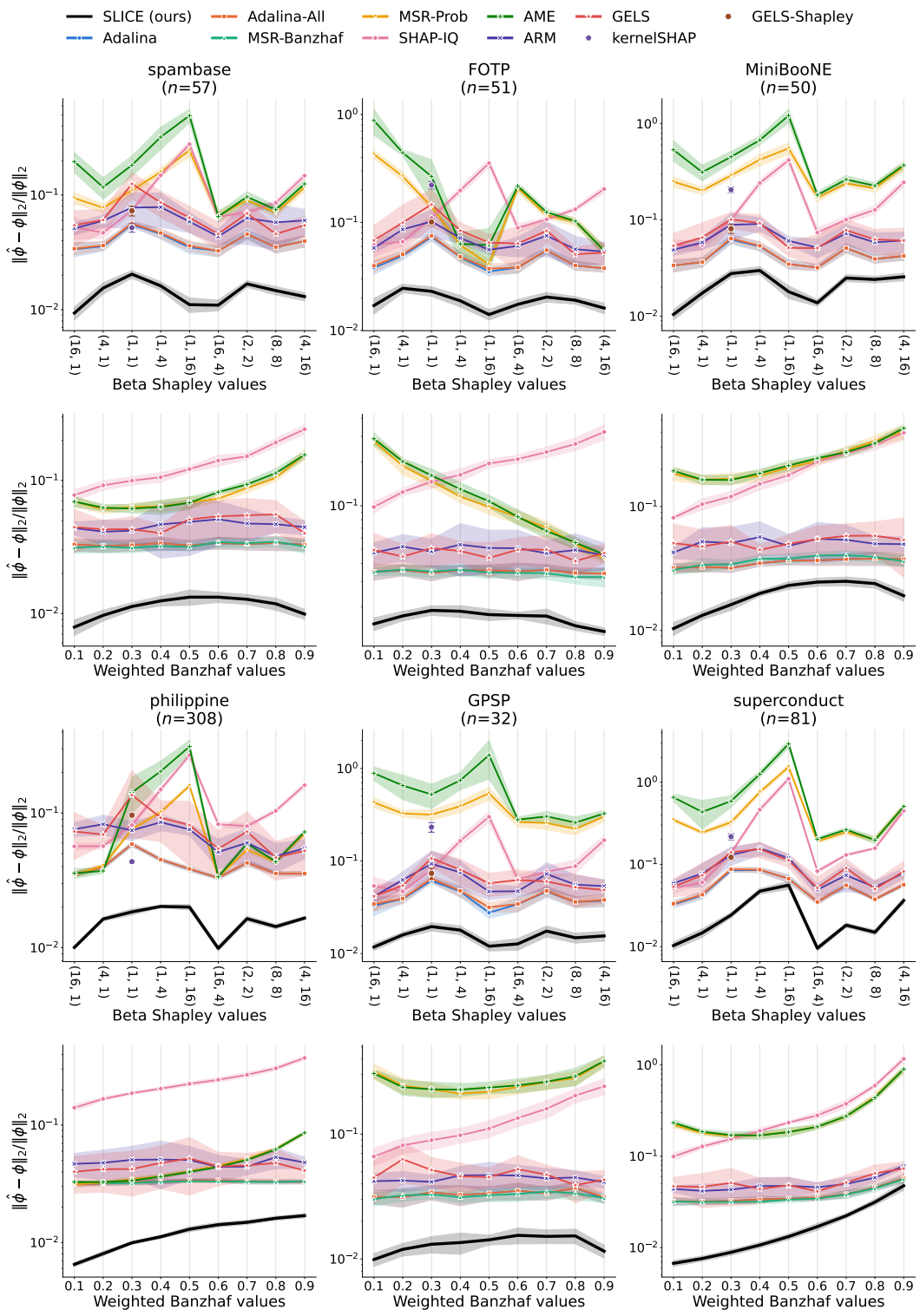

Figure 4: Relative ℓ_2 error across 6 datasets (spambase, FOTP, MiniBooNE, philippine, GPSP, superconduct). Top row: Beta Shapley results for varying (α, β) pairs of the Beta distribution. Bottom row: Weighted Banzhaf results for varying weight $w \in [0.1, 0.9]$. Values are mean $\pm$ 1 std over 10 random seeds.

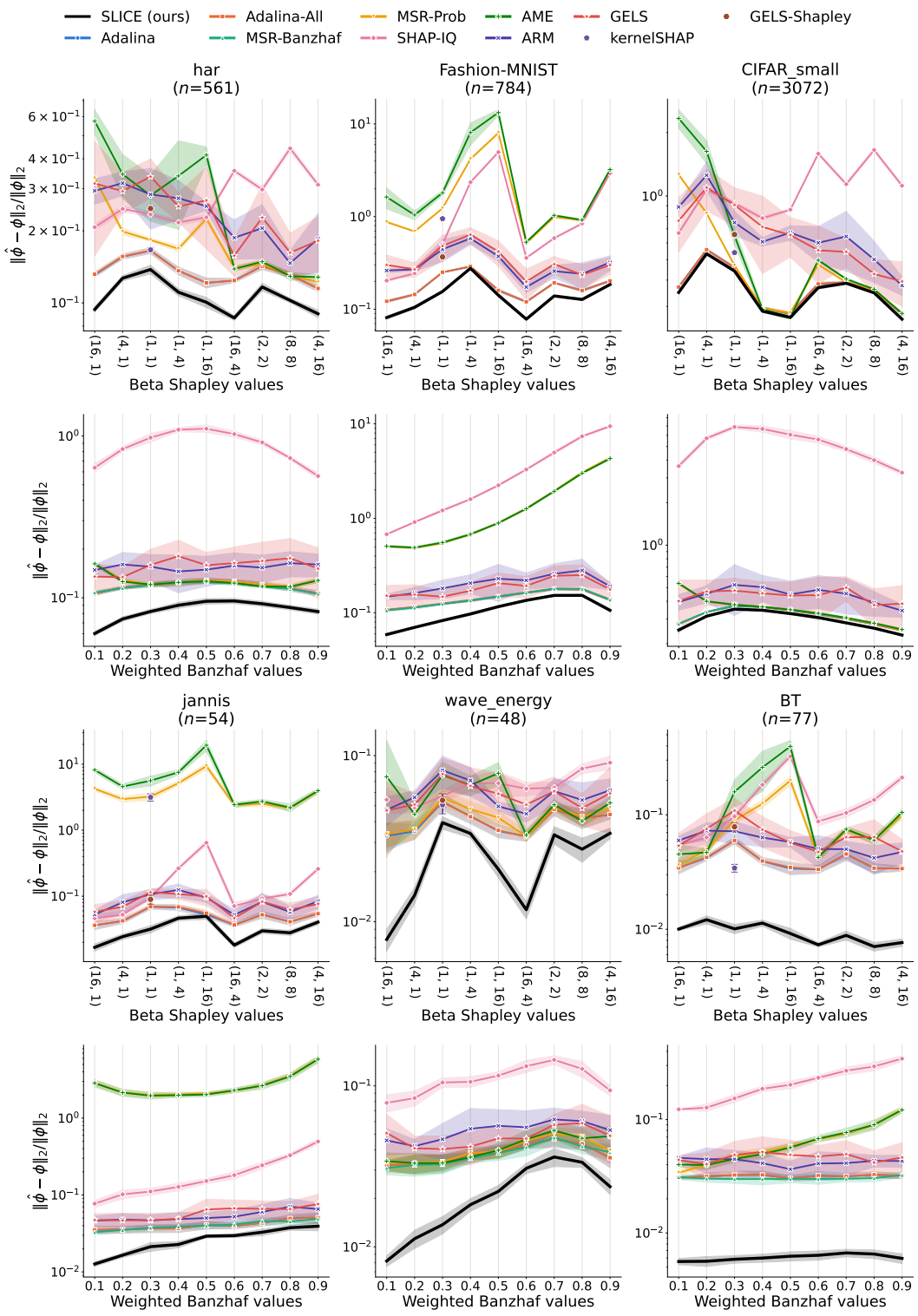

Figure 5: Relative ℓ_2 error across 6 datasets (har, Fashion-MNIST, CIFAR_small, jannis, wave_energy, BT). Top row: Beta Shapley results for varying (α, β) pairs of the Beta distribution. Bottom row: Weighted Banzhaf results for varying weight $w \in [0.1, 0.9]$. Values are mean $\pm$ 1 std over 10 random seeds.