ENDPOINT SENSITIVITIES GUIDE LIMITED-QUERY SHAPLEY REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Shapley attribution needs every coalition of input groups, and for an implicit graph neural network each coalition costs an equilibrium solve. An affordable budget can fit a polynomial surrogate's pairwise interactions but not its three-way ones, so one three-way term must be chosen before the sampled values can fit its coefficient. We let the explained model make that choice. Its endpoint sensitivities, measured only at the empty and the full coalition, select the term at the cost of two linear solves, and we prove that the coalition values cannot make the choice themselves, since their own endpoint information reduces to a single number beyond pairwise structure. Across four graph-classification datasets and twenty independently trained models, sensitivity-informed cubic Shapley regression cuts low-budget attribution error by over 75% against second-order PolySHAP, and matched controls that choose the term differently all have higher error, 1.55 times ours when our choice is permuted across groups. More broadly, model derivatives can choose where to spend scarce representational capacity, while the costly evaluations that define the target decide how much of it to use.

1 INTRODUCTION

The Shapley value of an input group is its expected marginal contribution to the model's output when the g groups arrive in uniformly random order (Shapley, 1953), so computing it exactly requires the model's value on every coalition, which keeps its own groups and replaces the others by a fixed baseline. A single coalition value can cost a retraining, a conditional inference or, for the implicit graph neural networks explained here, an equilibrium solve, so practical estimators sample a budget of coalitions and fit a surrogate (Lundberg & Lee, 2017; Covert & Lee, 2021). How much a budget buys depends on which coalitions are observed (Covert & Lee, 2021; Rundel et al., 2026), on which interactions the surrogate can represent (Fumagalli et al., 2026b;a) and on what structure the explained model exposes. Every interaction order the surrogate adds must be identified from the sample, and when coalitions are expensive the affordable budgets are exactly those that leave a high-order surrogate underdetermined.

At $g = 8$ groups this tension is concrete. Efficiency pins the fitted game to the observed full-coalition value and leaves a quadratic multilinear model 35 free coefficient directions and a cubic one 91. The confirmatory budgets of our primary study, $B \in \{48, 64, 80\}$, spend two evaluations on the empty and full coalitions and sample $m = B - 2 \in \{46, 62, 78\}$ interior ones, more than the quadratic count and fewer than the cubic one. No choice of interior coalitions then determines the cubic coefficient vector, so degree three, the first order with triple interactions, is also the first order the budget cannot resolve. Under the randomized complement-paired designs we study, even the Shapley value of a cubic game is out of reach in general, and no design the PolySHAP sampler drew at these budgets determines it (Appendix C). Structured designs with about g^2 chosen coalitions determine it for games of low interaction order (Hu & Wang, 2023; Zhou et al., 2026), while regression on randomized designs assumes nothing about the interaction order of the game.

Problem. *When the sampled coalitions cannot determine a complete cubic model, can information from the explained model select a useful cubic basis function before its coefficient is fitted?*

The Shapley value depends only on the coalition game, but the model that generates the game can report more than its coalition values. The models we explain accept fractional masks, so their games

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.

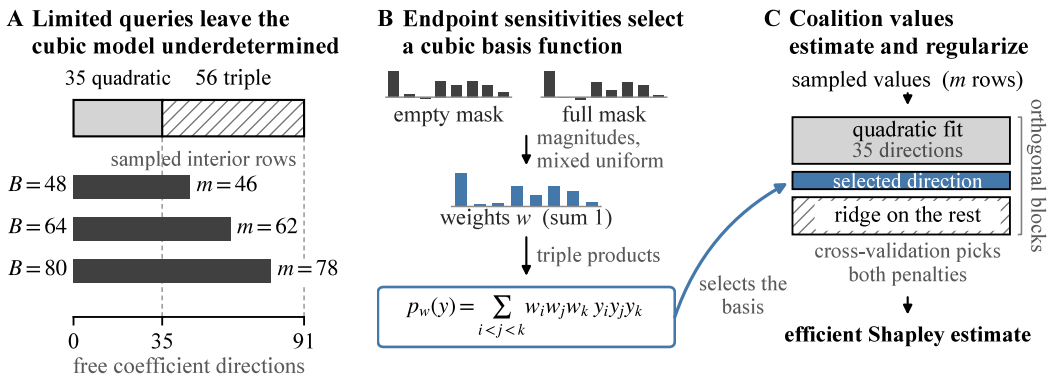


Figure 1: The estimator and its two information sources. **A:** the confirmatory budgets supply $m \in \{46, 62, 78\}$ interior coalition values, between the quadratic model's 35 free directions and the cubic model's 91. **B:** two linear solves at equilibria the budget already contains give the endpoint gradients, whose magnitudes set the weights w that select p_w (a held-out MUTAG target). **C:** the sampled values alone estimate all three blocks, and the exact Shapley values only evaluate the result.

extend differentiably in the mask, and the gradients of this extension at the empty and full masks have one coordinate per group, the endpoint sensitivities. Implicit differentiation delivers them at equilibria the budget already pays for (Gu et al., 2020), and the estimator reads them before any interior coalition, solely to choose a basis function. Their magnitudes, normalized to strictly positive weights w, select the cubic basis function

$$p_w(y) = e_3(w \odot y) = \sum_{i<j<k} w_i w_j w_k\, y_i y_j y_k,$$

the third elementary symmetric polynomial evaluated at the weighted coalition indicator $w \odot y$ with $y \in \{0, 1\}^g$ (Figure 1). The endpoint sensitivities thus choose the basis function, and the coalition values estimate its coefficient, regularize the remaining cubic space and produce the attribution.

1.1 Our Results

Our first result, Theorem 1, concerns the game itself. For $g \geq 3$, the values and gradients of the game's multilinear extension at the empty and full masks, taken modulo the degree-two games that the regression fits directly, retain exactly one scalar, and relabeling the groups leaves it unchanged. A rule that ignores quadratic additions and respects relabeling can therefore return only the line spanned by the cubic coefficients of e_3, which weights every triple equally, so an instance-specific choice needs information that distinguishes the groups. The model's gradients supply it because the game does not determine the model's mask extension, and distinct positive weights built from them select distinct lines once $g \geq 4$.

The estimator, sensitivity-informed cubic Shapley regression, adds the normalized basis function to the quadratic model as a single regressor q, which extends the surrogate beyond degree two at the cost of one coefficient. Theorem 2 reduces the joint efficiency-constrained fit to a scalar regression on the sampled values, and under a standard estimability condition the resulting attribution is exact whenever the target is a quadratic game plus a multiple of q. Generalized ridge regularizes the remaining cubic directions with separate penalties on the supplied coefficient and on the other directions, both selected by cross-validation on the sampled responses, and the target's exact Shapley vector never enters. Theorem 3 guarantees that the estimator reads no coalition value beyond the B supplied, returns an efficient attribution, recovers the same class exactly under matching conditions and commutes with relabeling of the groups.

The experiments isolate the mechanism with matched controls. The sharpest control keeps all eight weights and changes only which group each one belongs to, and in the predeclared four-dataset aggregate its error is 1.55 times the estimator's, with 95% hierarchical interval $[1.14, 4.37]$. Uniform weights, group-size weights and cubic ridge without a supplied regressor all have intervals above one, which locates the improvement in the direction the sensitivities select, and with any one

dataset removed all four point gains stay above 1.12. Against the quadratic baseline, second-order PolySHAP on the same paired design, the gain is 5.22. The ordering carries over to node-level explanations on three citation networks, where the groups have nearly equal sizes and group size therefore cannot explain the gain, and at equal Boolean coalition-query counts the estimator beats the reference implementation of OddSHAP in all 60 model-seed-budget cells.

1.2 Related Work

KernelSHAP casts Shapley estimation as weighted regression on sampled coalitions (Lundberg & Lee, 2017), and later work refines its sampling and variance (Covert & Lee, 2021; Musco & Witter, 2025) or chooses each next coalition by Bayesian experimental design (Rundel et al., 2026). These methods decide which rows the budget observes. Our estimator adds one column, fixed before any row is spent, so the two approaches compose.

PolySHAP extends the regression to interaction-informed polynomial surrogates (Fumagalli et al., 2026b), as k-additive approximations do for the Shapley value (Pelegrina et al., 2023; 2026), and our quadratic baseline is its second-order fit. OddSHAP shows that Shapley values depend only on a game's odd part and selects sparse odd interactions with a proxy trained on the sampled coalition values (Fumagalli et al., 2026a), and our direction is instead read from the model before sampling.

Interaction indices axiomatize how a game's higher-order structure is shared among its players (Grabisch & Roubens, 1999; Fujimoto et al., 2006), whereas Theorem 1 determines what a game's endpoint values and gradients reveal about it modulo quadratic games. Derivatives already drive several attribution methods. Integrated Gradients and Owen's diagonal formula integrate them along a path (Sundararajan et al., 2017; Owen, 1972), and ArchDetect detects interactions from mixed differences (Tsang et al., 2020). Our estimator uses them only to select the basis function.

The closest precedent is ControlSHAP, which uses the exact Shapley values of Taylor surrogates as control variates after the fit (Goldwasser & Hooker, 2024), and regression-adjusted Monte Carlo corrects a fitted surrogate similarly (Witter et al., 2025). Both act after the fit, and our estimator acts before it. MODEL SHAPLEY attributes to parameter blocks (Chu et al., 2025), a different estimand. Deep equilibrium models and implicit graph neural networks supply the model class and a standard differentiation calculus (Bai et al., 2019; Gu et al., 2020), and using the resulting gradients to select a regression basis is new. Explainers for graph neural networks optimize soft masks (Ying et al., 2019), score subgraphs by Shapley values (Yuan et al., 2021) or fit surrogates on perturbed inputs (Duval & Malliaros, 2021), and they assume cheap forward passes and bounded receptive fields, whereas for an implicit model one coalition value is an equilibrium solve whose mask response runs over all propagation lengths.

2 Endpoint Information in Coalition Games

Every budget pays for the empty and full masks, since efficiency needs them, and for the models we explain the derivatives there cost little more. Theorem 1 shows that for the game's own multilinear extension these endpoint values and gradients reduce, modulo degree-two games, to one scalar, which relabeling leaves unchanged and which, once $g \geq 4$, is compatible with more than one cubic direction. The estimator therefore reads the gradients of the model's own extension (Section 2.2).

2.1 Limited-Query Shapley Regression

Let $N = [g]$ index disjoint groups of input coordinates or nodes, with $g \geq 3$ throughout, let $\mathbf{1}_S \in \{0,1\}^g$ be the indicator of $S \subseteq N$, and write $\mathbf{1}$ for an all-ones vector of whatever length the context requires. A coalition S retains its groups and replaces the rest by a fixed baseline, and evaluating the explained model there and reading one scalar output defines the coalition game $v : 2^N \to \mathbb{R}$. Write $\bar{v}(S) = v(S) - v(\varnothing)$ and $d_v = \bar{v}(N)$. The Shapley value (Shapley, 1953) of group i is

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(g - |S| - 1)!}{g!} \big(v(S \cup \{i\}) - v(S)\big),$$

and an estimate $\widehat{\phi}$ is *efficient* when $\mathbf{1}^\top \widehat{\phi} = d_v$. Two of the B coalition evaluations go to the empty and full coalitions, which center the game and fix d_v, and the remaining $m = B - 2$ are the sampled interior coalitions.

Every game has a unique multilinear expansion

$$v(y) = \sum_{A \subseteq N} \mu_v(A) \prod_{i \in A} y_i, \tag{1}$$

whose coefficients are the Harsanyi dividends of v (Harsanyi, 1959). Let $\mathcal{G}_{\leq k}$ be the games whose expansion has degree at most k, the at most k-additive games of Grabisch (1997). Only $k \in \{2, 3\}$ occurs below. Reading Equation (1) on $[0,1]^g$ instead of on the Boolean cube gives the multilinear extension of v (Owen, 1972). Its homogeneous cubic coefficients are $\hom_3(v) = (\mu_v(T))_{|T|=3}$, and the symmetric cubic game $e_3(y) = \sum_{i<j<k} y_i y_j y_k$ has $\hom_3(e_3)$ all ones.

Polynomial Shapley regression fits a member of $\mathcal{G}_{\leq k}$ to the sampled values, whatever the interaction order of the game itself, and reads the attribution off the fitted coefficients, which takes at least as many equations as free coefficients. After the efficiency constraint, degree two leaves $g + \binom{g}{2} - 1$ free coefficients and degree three leaves $g + \binom{g}{2} + \binom{g}{3} - 1$, which for $g = 8$, the setting of every experiment below, are 35 and 91. Since $m = B - 2$,

$$m < 91 \quad \Longleftrightarrow \quad B < 93, \tag{2}$$

and our confirmatory budgets give $35 < m < 91$.

2.2 Differentiable Mask Extensions

The models we explain can also be evaluated on fractional masks, and this is the source of the information the estimator adds to the coalition values.

Definition 1. A *differentiable mask extension* of v is a continuously differentiable map f on an open set containing $[0,1]^g$ with $f(\mathbf{1}_S) = v(S)$ for every $S \subseteq N$. Its *first-order endpoint data* are

$$\mathcal{E}(f) = (v_0, v_1, K_0, K_1) = \big(f(0),\ f(\mathbf{1}),\ \nabla f(0),\ \nabla f(\mathbf{1})\big).$$

We call K_0 and K_1 the *endpoint gradients* and their coordinates the *endpoint sensitivities*.

One of a game's many differentiable mask extensions is intrinsic, its own multilinear extension, whose endpoint data $\mathcal{E}(v)$ are the subject of Theorem 1. The explained model supplies a second extension. Running it on a fractional mask $\theta \in [0,1]^g$, which interpolates each input of group i between its baseline and its observed value with weight θ_i (Appendix D), defines a differentiable mask extension f of v that the game does not determine, so two models can agree on every coalition while their endpoint gradients differ. The estimator reads this model extension at the two endpoints.

When the masked equilibrium $x^\star(\theta) = F(x^\star(\theta), \theta)$ comes from a continuously differentiable F that contracts in the state uniformly over an open set of masks containing $[0,1]^g$, the equilibrium is unique and differentiable in θ. The gradient of the scalar output at an endpoint then takes one adjoint solve with the operator $(I - \partial_x F)^\top$ and a single right-hand side, at an equilibrium the budget has already computed. The estimator thus consumes B coalition values plus two linear solves, which cost about as much as two coalition evaluations on our graph-classification models. Fixed-point architectures and their implicit differentiation are standard (Bai et al., 2019; Gu et al., 2020), and Appendix D gives the linear system, the contraction hypothesis with its verification, and the measured cost.

2.3 Endpoint Data Modulo Quadratic Games

Set the model aside and consider what a game's own extension determines. The regression fits every degree-two term directly, and adding a degree-two game leaves the cubic coefficients unchanged, so a rule aimed at those coefficients should read endpoint data modulo $\mathcal{G}_{\leq 2}$. Modulo $\mathcal{G}_{\leq 2}$ the endpoint data retain the single scalar

$$D(v_0, v_1, K_0, K_1) = v_1 - v_0 - \tfrac{1}{2} \sum_i \big(K_{0i} + K_{1i}\big). \tag{3}$$

Written in dividends (Appendix A), Equation (3) becomes $D(v) = \sum_{|A|\geq 3}\big(1 - \frac{|A|}{2}\big)\mu_v(A)$, so every dividend of degree at most two is annihilated, and up to scale D is the only linear functional of endpoint data with this property. We write $D(f)$ and $D(v)$ for its values at the two extensions above, and the two can differ.

Theorem 1 quantifies over maps that read endpoint data. An *endpoint rule* is a map Λ from first-order endpoint data to any fixed set. It is *quadratic-invariant* if $\Lambda(\mathcal{E}(v+b)) = \Lambda(\mathcal{E}(v))$ for all v and all $b \in \mathcal{G}_{\leq 2}$, and *equivariant* if $\Lambda(\mathcal{E}(\pi \cdot v)) = \pi \cdot \Lambda(\mathcal{E}(v))$ for every π in the symmetric group $\mathfrak{S}_g$, which acts by $(\pi \cdot v)(S) = v(\pi^{-1}(S))$ on games and by $(\pi \cdot c)_T = c_{\pi^{-1}(T)}$ on cubic coefficients.

Theorem 1 (Endpoint values and gradients modulo quadratic games)**.** *Let $g \geq 3$.*

(i) $D(v + b) = D(v)$ *for every game v and every $b \in \mathcal{G}_{\leq 2}$.*

(ii) For games v and v', $D(v) = D(v')$ if and only if $\mathcal{E}(v) - \mathcal{E}(v') = \mathcal{E}(b)$ for some $b \in \mathcal{G}_{\leq 2}$.

(iii) Every quadratic-invariant endpoint rule factors through D.

(iv) If $g \geq 4$, no quadratic-invariant endpoint rule returns $\mathrm{span}\{\mathrm{hom}_3(v)\}$ *for every game v with* $\mathrm{hom}_3(v) \neq 0$.

(v) A quadratic-invariant endpoint rule that is also equivariant and returns a nonzero line of cubic coefficients returns $\mathrm{span}\{\mathrm{hom}_3(e_3)\}$.

In words, a game's own endpoint data carry a single number beyond pairwise structure and cannot point to an instance-specific cubic direction. Appendix A proves it. A degree-two game with prescribed pair-coefficient row sums, which exists exactly because $g \geq 3$, realizes any endpoint data with $D = 0$, and part (iv) is witnessed by $y_1y_2y_3$ and $y_1y_2y_4$, which share D but span different lines. The model's endpoint gradients are not fixed by the game, so they carry information beyond what Theorem 1 classifies, and Section 3 builds the cubic line from them.

3 From Endpoint Sensitivities to a Cubic Basis Function

The sample has room beyond the quadratic model but not for the full cubic one, so the estimator spends one degree of freedom on a single cubic direction chosen from the endpoint gradients, whose magnitudes become strictly positive weights and then a basis function built from their triple products.

3.1 Sensitivity Weights

We reduce each group's two sensitivities to one magnitude, normalize, and mix with the uniform vector,

$$\kappa_i = \sqrt{K_{0i}^2 + K_{1i}^2}, \qquad w_i = (1-\varepsilon)\frac{\kappa_i}{\sum_j \kappa_j} + \frac{\varepsilon}{g}, \tag{4}$$

with $\varepsilon = 0.05$ fixed before the estimator was evaluated and with $w_i = 1/g$ when every κ_j vanishes. We use the magnitude because a signed combination cancels when a group responds strongly but oppositely at the two endpoints, whereas κ_i vanishes only when both sensitivities do. The weights are strictly positive, sum to one, relabel with the groups and ignore readout translation and rescaling, and the rule reads only K_0 and K_1, never a coalition value or $\phi(v)$.

3.2 Weighted Cubic Basis

The weight of a triple is the product of its members' weights, so the basis function is

$$p_w(y) = e_3(w \odot y) = \sum_{i<j<k} w_i w_j w_k \, y_i y_j y_k, \tag{5}$$

where $w \odot y$ is the coordinatewise product. Write $\xi(y) = w^\top y$ for the weighted coalition size. Expanding ξ^3 on the Boolean cube gives $y_iy_jy_k$ the coefficient $3!\, w_iw_jw_k$, since a repeated index drops below degree three when y_i^2 becomes y_i. The multiplicative triple weighting is therefore the

degree-three content of ξ^3 up to that factor. It puts cubic capacity on triples of high-sensitivity groups and costs the regression one coefficient instead of $\binom{g}{3}$.

The regression uses the unit-norm representative $q = p_w / \|\hom_3(p_w)\|_2$, well defined because the weights are strictly positive. Section 4 needs the exact Shapley vector of q, and it has a closed form. A monomial's dividend splits equally among its players (Appendix B), so summing over the triples that contain i gives

$$\phi_i(p_w) = \frac{w_i}{3} \sum_{\substack{j<k \\ j,k\neq i}} w_j w_k, \tag{6}$$

and $\phi(q) = \phi(p_w)/\|\hom_3(p_w)\|_2$ by linearity.

The estimator depends on q only through the line $\operatorname{span}\{q\}$ (Appendix C), and for $g \geq 4$ distinct strictly positive normalized weight vectors give distinct lines (Appendix B). Uniform weights give $p_w = g^{-3} e_3$ for every target, whose cubic coefficients span the symmetric line of part (v) of Theorem 1, and Section 5 runs this choice as a control.

Hermite interpretation. The line of p_w also carries the cubic part of an interpolant of the model's endpoint data. Appendix B constructs a cubic h that matches all of $\mathcal{E}(f)$ by adding the Hermite blending term $(3\xi^2 - 2\xi^3)D(f)$ to a quadratic, so its only degree-three term is $-2D(f)\,\xi^3$. Writing $\mathrm{ml}(h)$ for the multilinear polynomial that agrees with h on $\{0,1\}^g$,

$$\hom_3(\mathrm{ml}(h)) = -12\, D(f)\ \hom_3(p_w), \tag{7}$$

by the expansion of ξ^3 above. The weights thus fix the line and $-12D(f)$ the multiple along it. The estimator keeps only the line and estimates the coefficient from the sampled values, so q is supplied even when $D(f) = 0$.

4 Sensitivity-Informed Cubic Shapley Regression

With q fixed, it remains to fit its coefficient and to decide how much residual cubic space to keep, both under the efficiency constraint, which is affine and shifts the origin of every fit and residual. Theorem 2 handles the first step and generalized ridge the second.

4.1 Efficiency-Constrained Quadratic Regression

Write $\Omega = (S_1, \ldots, S_m)$ for the sampled interior coalitions. For $k \in \{2, 3\}$, let $x_k(S)$ list the nonconstant Boolean monomials of degree at most k, let X_k have rows $x_k(S_j)^\top$, and let $W \succ 0$ be the diagonal square root of the row weights that the PolySHAP sampler returns (Appendix E). For any game z put $\bar{z}(S) = z(S) - z(\varnothing)$ and $d_z = \bar{z}(N)$, with $\bar{z}_\Omega$ the vector of $\bar{z}$ on Ω. Every monomial equals one at the full coalition, so fitting z by a member of $\mathcal{G}_{\leq k}$ under efficiency means

$$\min_a \|W(X_k a - \bar{z}_\Omega)\|_2^2 \quad \text{subject to} \quad \mathbf{1}^\top a = d_z. \tag{8}$$

Feasible coefficient vectors differ by directions orthogonal to $\mathbf{1}$. Write P_k for the orthogonal projector onto those directions and $A_k = W X_k P_k$ for the sampled response they produce, and let $C_k z$ be the minimum-norm solution of Equation (8), which is linear in z and defined for every sampled design. Pinning down this solution matters because minimizers with the same residual can differ in attribution.

Let Φ_k be the exact Shapley map on coefficients, which divides each monomial coefficient equally among the players of its monomial, and put $\mathcal{A}_k z = \Phi_k C_k z$. Summing a coefficient's shares returns the coefficient, so $\mathbf{1}^\top \Phi_k = \mathbf{1}^\top$ and $\mathcal{A}_k z$ is efficient. The sampled residual of the constrained degree-two fit, $\rho_2(z) = W\big(\bar{z}_\Omega - X_2 C_2 z\big)$, inherits the linearity of C_2.

Exact recovery in Theorem 2 asks $\mathcal{A}_2$ to reproduce ϕ on $\mathcal{G}_{\leq 2}$, which holds under the standard estimability condition

$$\ker(A_2) \subseteq \ker(\Phi_2 P_2), \tag{9}$$

stating that coefficient perturbations invisible to the sampled design leave the attribution unchanged. Full projected rank, 35 at $g = 8$, implies it, but the theorems need only the condition, which

holds in every design we drew even where pairing caps the rank at 30 (Appendices C and E). With complement-paired sampling $\mathcal{A}_2$ is second-order PolySHAP, which coincides with paired KernelSHAP whenever X_2 has full column rank (Fumagalli et al., 2026b).

4.2 Constrained Partial Regression

Supplying q means fitting v by a member of $\mathcal{G}_{\leq 2}$ together with a multiple of q. Write $r = \rho_2(v)$ and $u = \rho_2(q)$ for the sampled residuals of target and regressor, $\delta = \|u\|_2^2$ for the energy of u, and $\Delta_q = \phi(q) - \mathcal{A}_2 q$ for the attribution gained by replacing the quadratic fit of q with its exact Shapley vector. For $\tau \geq 0$ the augmented problem is

$$\begin{aligned} &\min_{a,\beta} \ \|W(X_2 a + \beta \bar{q}_\Omega - \bar{v}_\Omega)\|_2^2 + \tau\beta^2 \\ &\text{subject to } \mathbf{1}^\top a + \beta d_q = d_v, \end{aligned} \tag{10}$$

with a selected by the same minimum-norm rule. Profiling out the quadratic block reduces Equation (10) to one dimension, which is the classical partialling-out identity (Frisch & Waugh, 1933; Lovell, 1963) carried through an affine constraint and a possibly singular design.

Theorem 2 (Constrained partial regression along a supplied cubic regressor)**.** *Let $\tau \geq 0$ with $\delta + \tau > 0$. The coefficient minimizing Equation (10) and the attribution read off its solution are*

$$\widehat{\beta}_\tau = \frac{u^\top r}{\delta + \tau}, \qquad \widehat{\phi}_q(v) = \mathcal{A}_2 v + \widehat{\beta}_\tau \Delta_q, \tag{11}$$

and the attribution is efficient for every τ. If in addition $\tau = 0$, $\delta > 0$, Equation (9) holds, and $v = b + \beta_\star q$ with $b \in \mathcal{G}_{\leq 2}$, then $\widehat{\beta}_0 = \beta_\star$ and $\widehat{\phi}_q(v) = \phi(v)$.

Proof sketch. For fixed β, Equation (10) is Equation (8) for $v - \beta q$, whose constrained residual is $r - \beta u$ by linearity, so the profiled objective $\|r - \beta u\|_2^2 + \tau\beta^2$ gives $\widehat{\beta}_\tau$ and linearity of $\mathcal{A}_2$ gives the attribution, which is efficient because $\phi(q)$ and $\mathcal{A}_2 q$ both sum to d_q. Appendix C completes the proof. □

4.3 Generalized Ridge on the Residualized Cubic Space

The estimator keeps whatever cubic directions the quadratic model and q leave in the sample. When $\delta > 0$, let $H_2 = A_2 A_2^\dagger$ project onto the sampled quadratic space, with $\dagger$ the Moore–Penrose pseudoinverse, and let $P_u = uu^\top/\delta$ project onto the residualized regressor. Since u is a degree-two residual, $\mathrm{range}(A_2)$, $\mathrm{span}\{u\}$ and the residualized cubic space $\mathrm{range}(B_q)$, with $B_q = (I - H_2 - P_u)A_3$, are mutually orthogonal, and each column of B_q is the sampled response of one cubic direction with the quadratic model and u removed. Let θ_λ be the minimum-norm ridge fit of r on B_q with penalty λ (Hoerl & Kennard, 1970), and let

$$\mathcal{D}_q = \Phi_3 P_3 - \Phi_2 P_2 A_2^\dagger A_3 - \Delta_q \frac{u^\top A_3}{\delta} \tag{12}$$

send a cubic coefficient direction to the attribution that survives removing its quadratic refit and its component along q. A candidate estimate keeps the quadratic fit, shrinks the coefficient of q by $\gamma \in [0, 1]$ and adds the ridge correction,

$$\widetilde{\phi}_{\gamma,\lambda} = \mathcal{A}_2 v + \gamma \widehat{\beta}_0 \Delta_q + \mathcal{D}_q \theta_\lambda. \tag{13}$$

Since $\widehat{\beta}_\tau = \gamma \widehat{\beta}_0$ for $\gamma = \delta/(\delta + \tau)$ by Equation (11), both parameters are ridge penalties on orthogonal subspaces. Since $\mathbf{1}^\top \Phi_k = \mathbf{1}^\top$, $\mathbf{1}^\top P_k = 0$ and $\mathbf{1}^\top \Delta_q = 0$, the vector $\mathbf{1}^\top$ annihilates every term of Equation (12), so both corrections sum to zero and every candidate is efficient.

We call a pair (γ, λ) *admissible* if it leaves positive residual degrees of freedom. Generalized cross-validation (Golub et al., 1979) scores a fixed grid of admissible pairs, and a positive-part shrinkage factor in $[0, 1]$ scales the selected correction (Appendix C). A *retention test* drops the supplied block when δ is negligible against the sampled norm of q, that is, when q is numerically aliased with the quadratic model, and the estimator then runs ridge on $(I - H_2)A_3$ only, the branch that the cubic-ridge control of Section 5 runs on every target.

4.4 Algorithm and Guarantees

The algorithm. Only the first of three stages reads the model. Selection turns K_0 and K_1 into w and q. Fitting then computes $\mathcal{A}_2 v$, r and u and applies the retention test, and regularization scores the candidate grid on the same sampled responses and scales the selected correction. No stage reads $\phi(v)$, and Appendix C gives the pseudocode.

Theorem 3 (Estimator guarantees)**.** *Suppose $q \in \mathcal{G}_{\leq 3}$, and suppose the branch chosen by the retention test has an admissible candidate in its grid.*

(i) The estimator uses no coalition value beyond the B supplied as input.

(ii) Every admissible candidate, the selected candidate and the scaled output are efficient.

(iii) If the supplied block is retained, Equation (9) holds, an admissible candidate with $\gamma = 1$ exists, and $v \in \mathcal{G}_{\leq 2} \oplus \mathrm{span}\{q\}$, then the selected candidate and the output both equal $\phi(v)$.

(iv) Relabeling the players relabels the output by the same permutation.

Appendix C proves the theorem, whose substance is part (iii). If $v = b + \beta_\star q$, then $r = \beta_\star u$, which $B_q^{\dagger}$ annihilates, so every candidate with $\gamma = 1$ fits the residual exactly. When $\beta_\star \neq 0$ only these candidates score zero, and their zero residual sets the scaling factor to one. When $\beta_\star = 0$ or $\Delta_q = 0$ the quadratic estimate is already exact.

5 Experiments

Supplying a cubic regressor adds one degree of freedom and lets the endpoint sensitivities choose where it points. Three controls change where it points, a fourth withholds it, and two further studies change the explanation setting and add a contemporary limited-query baseline.

5.1 Questions and Protocol

The primary study covers MUTAG, PROTEINS, NCI1 and IMDB-BINARY, with five independently trained implicit graph neural networks each and 24 held-out graphs per model, each partitioned into eight graph-distance Voronoi node groups. Each budget $B \in \{48, 64, 80\}$ runs 50 paired mask designs against the full-input predicted-class logit. At $g = 8$ the complete game is affordable once per unit, which yields the exact Shapley value for measuring error, and the estimator never sees it. Appendix E gives the splits, architectures, training and grouping rules, and Appendix G the artifact[1]. The confirmatory analysis is restricted to the regime in which no design can identify the cubic coefficient vector because every design has fewer than 91 interior rows, which by Equation (2) means $B < 93$, and of the eight budgets in our sweep (Appendix E) exactly these three satisfy it.

With $\mathrm{SSE}_{d,b,a}$ the squared attribution error of algorithm a on dataset d at budget b, summed over seeds, graphs, repetitions and coordinates, and quad the quadratic estimator, the dataset-budget-balanced relative sum of squared errors and the gain of A over R are

$$T_a = \sum_d \sum_b \frac{\mathrm{SSE}_{d,b,a}}{\mathrm{SSE}_{d,b,\mathrm{quad}}}, \qquad G(A/R) = \frac{T_R}{T_A}. \tag{14}$$

Dividing by the quadratic error puts the twelve cells on a common scale, so $T_{\mathrm{quad}} = 12$ and a gain above one favors A.

Every in-house control receives the same masks, coalition values, sampling normalization and target columns, and intervals come from a 100,000-draw hierarchical bootstrap over datasets, model seeds, graphs and paired repetitions (Appendix E). The four controls, ε, the candidate grid and the bootstrap were all fixed before any error was computed.

5.2 Primary Mechanism Comparisons

Three of the four controls run the identical estimator on the identical coalition values and change only the weight vector in Equation (5), to $w_i = 1/g$, to Equation (4) built from each group's node

[1] anonymous.4open.science/r/endpoint-sensitivities-shapley-artifact

count instead of its sensitivity magnitude, or to permuted sensitivity weights. The permutation control averages the squared error over all 105 fixed-point-free involutions of the groups, a complete conjugacy class, so it is invariant to relabeling and keeps the multiset of weights while detaching each weight from its group. The fourth control drops the supplied block and ridge-regresses on the whole cubic space after quadratic residualization.

Figure 2 collects the four comparisons, and every interval lies above one. The permutation control is the sharpest, because it changes only which group each weight belongs to, and its gain is 1.55 with interval $[1.14, 4.37]$. Uniform weights (1.29, $[1.02, 3.01]$) and group-size weights (1.17, $[1.04, 1.37]$) show that a per-target choice beats a single symmetric term and that the sensitivities carry information that group size, the most direct structural proxy for influence, does not. Cubic ridge without a supplied block (1.35, $[1.06, 3.33]$) shows that the selected direction adds to what cubic regularization achieves alone. Removing any one dataset leaves every point gain above 1.12.

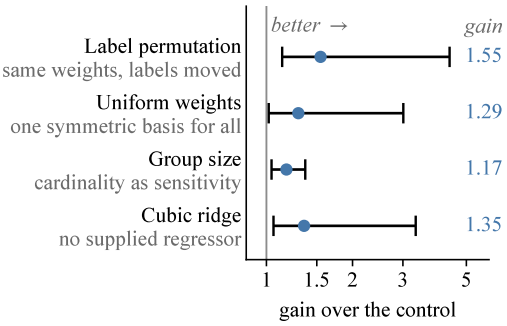


Figure 2: Gains T_R/T_A (Equation (14)) over the four matched controls in the predeclared low-budget aggregate, with 95% hierarchical bootstrap intervals, on a logarithmic scale.

The sensitivity weights change the direction of the supplied regressor in the sample. The retention test kept it in all 72,000 low-budget target-design pairs, and the sine of the angle between the residualized regressor and its uniform-weight counterpart has median 0.35 (Appendix F). Against the quadratic estimator the whole cubic procedure gains 5.22.

5.3 Balanced Groups and a Contemporary Baseline

A node-level study explains the predicted-class logits of individual nodes of Cora, CiteSeer and PubMed (Sen et al., 2008; Yang et al., 2016) under trained implicit graph convolutional models, with its own protocol of nine main configurations whose eight feature groups differ in size by at most one, so group size cannot explain a gain. The quadratic estimator, cubic ridge without a supplied regressor and the uniform-weight control all have higher pooled error there, by ratios of up to 26.95.

The final comparison is with OddSHAP (Fumagalli et al., 2026a), run from its authors' implementation on the same datasets and budgets with each method's native sampler, so it equalizes Boolean coalition queries and our estimator additionally spends its two linear solves. The proposed estimator has lower error in all 12 dataset-budget and all 60 model-seed-budget cells, and balancing the twelve cells as in Equation (14) with OddSHAP as the reference gives a gain of 21.63.

6 Discussion

At high order, polynomial Shapley regression is limited as much by the coefficient directions a budget must resolve as by the number of coalition values it can afford, and endpoint sensitivities fixed before any interior coalition value is read can choose the direction. Theorem 1 shows that a game's own endpoint data reduce, modulo quadratic games, to one scalar that relabeling leaves unchanged, and Theorem 2 reduces the fit along any externally chosen direction to a scalar regression on the sampled values. Where fewer interior rows are sampled than the cubic model has free directions, the matched controls trace the improvement to the direction the sensitivities choose.

Any differentiable network, deep equilibrium model or differentiable optimization layer supplies a mask extension and endpoint sensitivities computed without reading the sampled responses, and any polynomial family with closed-form Shapley vectors supplies the basis functions. The approach pays off most where coalition values are expensive, as for the implicit models studied here, where two linear solves choose the cubic term and the coalition values that define the target estimate its coefficient.

AI USE STATEMENT

The authors used OpenAI's ChatGPT and Anthropic's Claude to help organize and polish the exposition of this paper, based on the authors' own drafts. The same tools assisted in surveying the related literature, in understanding the current state of research in this area and its gaps, and in flagging possible errors in draft proofs. Within an experimental framework of the authors' own design, the tools were further used to implement the methods of this paper in code and to help produce the figures. The authors formulated and proved all mathematical claims in this paper, verified and tested all experimental code, and checked, revised, and finalized every statement, proof, figure, and reference. They take full responsibility for the content of the submission.

REPRODUCIBILITY STATEMENT

The proofs are in Appendix A for Theorem 1, in Appendix B for the constructions of Section 3, and in Appendix C for Theorems 2 and 3, and Appendix D gives the implicit differentiation behind the endpoint sensitivities and their cost. The estimator is specified in Algorithm 1, with its candidate grid, tie-breaking rule and retention threshold in Appendix C. Appendix E specifies the datasets, models, groupings, budgets, controls and the predeclared estimand of Section 5, and Appendix F reports the complete results. The artifact at `https://anonymous.4open.science/r/endpoint-sensitivities-shapley-artifact` contains the frozen results of every experiment, the pinned implementations of PolySHAP and OddSHAP, and the script `reproduce.py`, which verifies the frozen files and rebuilds the reported tables. Appendix G maps its identifiers to the names used in the paper.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems 32*, 2019.

Xu Chu, Xinke Jiang, Rihong Qiu, Jiaran Gao, and Junfeng Zhao. MODEL SHAPLEY: Find your ideal parameter player via one gradient backpropagation. In *Advances in Neural Information Processing Systems 38*, 2025.

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, 2021.

Alexandre Duval and Fragkiskos D. Malliaros. GraphSVX: Shapley value explanations for graph neural networks. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2021)*, volume 12976 of *Lecture Notes in Computer Science*. Springer, 2021. doi: 10.1007/978-3-030-86520-7_19.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. doi: 10.2307/1907330.

Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55 (1):72–99, 2006. doi: 10.1016/j.geb.2005.03.002.

Fabian Fumagalli, Landon Butler, Justin Singh Kang, Kannan Ramchandran, and R. Teal Witter. An odd estimator for Shapley values. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. Accepted at ICML 2026. arXiv:2602.01399.

Fabian Fumagalli, R. Teal Witter, and Christopher Musco. PolySHAP: Extending KernelSHAP with interaction-informed polynomial regression. In *Proceedings of the 14th International Conference on Learning Representations*, 2026b. arXiv:2601.18608.

Jeremy Goldwasser and Giles Hooker. Stabilizing estimates of Shapley values with control variates. In *Explainable Artificial Intelligence (xAI 2024)*, volume 2154 of *Communications in Computer and Information Science*, pp. 416–439. Springer, 2024. doi: 10.1007/978-3-031-63797-1_21.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. doi: 10.1080/00401706.1979.10489751.

Michel Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997. doi: 10.1016/S0165-0114(97)00168-1.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999. doi: 10.1007/s001820050125.

Fangda Gu, Heng Chang, Wenwu Zhu, Somayeh Sojoudi, and Laurent El Ghaoui. Implicit graph neural networks. In *Advances in Neural Information Processing Systems 33*, 2020.

John C. Harsanyi. A bargaining model for the cooperative n-person game. In *Contributions to the Theory of Games, Volume IV*, number 40 in Annals of Mathematics Studies. Princeton University Press, 1959. doi: 10.1515/9781400882168-019.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Linwei Hu and Ke Wang. Computing SHAP efficiently using model structure information, 2023. arXiv:2309.02417.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963. doi: 10.1080/01621459.1963.10480682.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 2017.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020. arXiv:2007.08663.

Christopher Musco and R. Teal Witter. Provably accurate Shapley value estimation via leverage score sampling. In *International Conference on Learning Representations (ICLR)*, pp. 64532–64559, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a224ff18cc99a71751aa2b79118604da-Abstract-Conference.html.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5-Part-2):64–79, 1972. doi: 10.1287/mnsc.18.5.64.

Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, and Michel Grabisch. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artificial Intelligence*, 325:104014, 2023. doi: 10.1016/j.artint.2023.104014.

Guilherme Dean Pelegrina, Patrick Kolpaczki, and Eyke Hüllermeier. Shapley value approximation based on k-additive games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17197–17205, 2026. doi: 10.1609/aaai.v40i20.38770.

David Rundel, Fabian Fumagalli, Maximilian Muschalik, Bernd Bischl, and Matthias Feurer. ShaplEIG: Bayesian experimental design for Shapley value estimation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. Accepted at ICML 2026. arXiv:2606.02247.

Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: 10.1609/aimag.v29i3.2157.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, Volume II*, number 28 in Annals of Mathematics Studies. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? Interpretable attribution for feature interactions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6147–6159. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/443dec3062d0286986e21dc0631734c9-Abstract.html.

R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted Monte Carlo estimators for Shapley values and probabilistic values. In *Advances in Neural Information Processing Systems*, volume 38, pp. 24027–24058, 2025. doi: 10.52202/085713-0715.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48, 2016.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems 32*, 2019. arXiv:1903.03894.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12241–12252, 2021.

Zheng Zhou, Robert Mee, Herbert Hamers, and Wei Zheng. Fast approximation of Shapley values through fractional factorial designs. *Journal of the American Statistical Association*, 121(553): 525–535, 2026. doi: 10.1080/01621459.2025.2529027.

NOTATION

The appendices use the notation of the main text, collected here for reference.

Let $N = [g]$ index the player groups, write $\mathbf{1}_S \in \{0,1\}^g$ for the indicator of $S \subseteq N$, and write $\mathbf{1}$ for an all-ones vector of whatever length its context requires. A coalition game is a map $v : 2^N \to \mathbb{R}$, and the Shapley value of group i is

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(g - |S| - 1)!}{g!} \big(v(S \cup \{i\}) - v(S)\big).$$

It is efficient, that is, $\sum_i \phi_i(v) = v(N) - v(\varnothing)$. Every game has a unique multilinear expansion $v(y) = \sum_{A \subseteq N} \mu_v(A) \prod_{i \in A} y_i$, whose coefficients are its Harsanyi dividends, and $\mathcal{G}_{\leq k}$ denotes the games whose expansion has degree at most k. Reading that expansion on $[0,1]^g$ gives the multilinear extension of v. Its homogeneous cubic coefficients are $\hom_3(v) = (\mu_v(T))_{|T|=3}$, and $e_3(y) = \sum_{i<j<k} y_i y_j y_k$ is the symmetric cubic game.

A *differentiable mask extension* of v is a continuously differentiable map f on an open set containing $[0,1]^g$ with $f(\mathbf{1}_S) = v(S)$ for every $S \subseteq N$. Its *first-order endpoint data* are

$$\mathcal{E}(f) = (v_0, v_1, K_0, K_1) = \big(f(0),\ f(\mathbf{1}),\ \nabla f(0),\ \nabla f(\mathbf{1})\big),$$

and K_0, K_1 are the *endpoint gradients*, their player-indexed coordinates the *endpoint sensitivities*. Put

$$D(v_0, v_1, K_0, K_1) = v_1 - v_0 - \tfrac{1}{2} \sum_i \big(K_{0i} + K_{1i}\big).$$

We write $D(f)$ for this functional evaluated at $\mathcal{E}(f)$, and $\mathcal{E}(v)$ and $D(v)$ for the same objects at the game's own multilinear extension.

Throughout, $g \geq 3$. For one scalar target the endpoint sensitivities define weights

$$\kappa_i = \sqrt{K_{0i}^2 + K_{1i}^2}, \qquad w_i = (1 - \varepsilon)\frac{\kappa_i}{\sum_j \kappa_j} + \frac{\varepsilon}{g}$$

with $\varepsilon = 0.05$, and $w_i = 1/g$ when $\sum_j \kappa_j = 0$. The weights are strictly positive and sum to one. They select the cubic basis function and its normalized form,

$$p_w(y) = e_3(w \odot y) = \sum_{i<j<k} w_i w_j w_k\, y_i y_j y_k,$$

$$q = p_w / \left\|\hom_3(p_w)\right\|_2,$$

where $w \odot y$ is the coordinatewise product and q is the regressor supplied to the coalition regression.

The regression runs on a fixed design. Let $\Omega = (S_1, \ldots, S_m)$ be the sampled interior coalitions of a budget $B = m + 2$, let $x_k(S)$ list the nonconstant Boolean monomials of degree at most k, let X_k have rows $x_k(S_j)^\top$, and let $W \succ 0$ be the diagonal square root of the sampling weights. Fitting a game z by a member of $\mathcal{G}_{\leq k}$ under efficiency means

$$\min_a \|W(X_k a - \bar{z}_\Omega)\|_2^2 \quad \text{subject to} \quad \mathbf{1}^\top a = d_z,$$

where $\bar{z}(S) = z(S) - z(\varnothing)$ and $d_z = \bar{z}(N)$. Write P_k for the orthogonal projector onto $\mathbf{1}^\perp$, put $A_k = W X_k P_k$, and let $C_k z$ be the minimum-norm solution. With Φ_k the map dividing each monomial coefficient equally among the players of its monomial, $\mathcal{A}_k z = \Phi_k C_k z$ and $\mathbf{1}^\top \Phi_k = \mathbf{1}^\top$. The constrained degree-two residual is $\rho_2(z) = W(\bar{z}_\Omega - X_2 C_2 z)$, and

$$r = \rho_2(v), \qquad u = \rho_2(q),$$

$$\delta = \|u\|_2^2, \quad \Delta_q = \phi(q) - \mathcal{A}_2 q.$$

Throughout, $M^\dagger$ is the Moore–Penrose pseudoinverse of M, $\|\cdot\|_F$ the Frobenius norm, $(a)_+ = \max\{a, 0\}$, and $\mathfrak{S}_k$ the symmetric group on k letters.

A Endpoint Information Modulo Quadratic Games

This appendix proves the endpoint theorem of the main paper and shows that the condition $g \geq 4$ in part (iv) is sharp.

Throughout, $g \geq 3$. An *endpoint rule* is a map Λ from first-order endpoint data to a fixed set of outputs. It is *quadratic-invariant* if $\Lambda(\mathcal{E}(v+b)) = \Lambda(\mathcal{E}(v))$ for every game v and every $b \in \mathcal{G}_{\leq 2}$, and *equivariant* if $\Lambda(\mathcal{E}(\pi \cdot v)) = \pi \cdot \Lambda(\mathcal{E}(v))$ for every $\pi \in \mathfrak{S}_g$, where $(\pi \cdot v)(S) = v(\pi^{-1}(S))$ and $\mathfrak{S}_g$ acts on cubic coefficient vectors by $(\pi \cdot c)_T = c_{\pi^{-1}(T)}$.

Theorem A.1. *Let $g \geq 3$.*

(i) $D(v+b) = D(v)$ for every game v and every $b \in \mathcal{G}_{\leq 2}$.

(ii) For games v and v', $D(v) = D(v')$ if and only if $\mathcal{E}(v) - \mathcal{E}(v') = \mathcal{E}(b)$ for some $b \in \mathcal{G}_{\leq 2}$.

(iii) Every quadratic-invariant endpoint rule factors through D.

(iv) If $g \geq 4$, no quadratic-invariant endpoint rule returns $\operatorname{span}\{\hom_3(v)\}$ for every game v with $\hom_3(v) \neq 0$.

(v) A quadratic-invariant endpoint rule that is also equivariant and returns a nonzero line of cubic coefficients returns $\operatorname{span}\{\hom_3(e_3)\}$.

A.1 D on Dividends

Differentiating the multilinear expansion gives $K_{0i} = \mu_v(\{i\})$, since at $y = 0$ only the singleton term survives, and $K_{1i} = \sum_{A \ni i} \mu_v(A)$, since at $y = \mathbf{1}$ every set containing i contributes its dividend. Hence $\sum_i K_{0i} = \sum_{|A|=1} \mu_v(A)$ and $\sum_i K_{1i} = \sum_A |A|\ \mu_v(A)$, while $v(N) - v(\varnothing) = \sum_{A \neq \varnothing} \mu_v(A)$. A set A with $|A| \geq 2$ therefore enters D with coefficient $1 - |A|/2$, a singleton with $1 - \frac{1}{2}(1+1) = 0$, and the empty set with zero. The coefficient vanishes for $|A| \leq 2$, leaving

$$D(v) = \sum_{|A| \geq 3} \Big(1 - \tfrac{|A|}{2}\Big) \mu_v(A). \tag{A.1}$$

Remark 1. D has a second reading, in terms of the extension rather than the game. For a differentiable mask extension f put $\psi(s) = f(s\mathbf{1})$, so that $\psi'(0) = \sum_i K_{0i}$ and $\psi'(1) = \sum_i K_{1i}$. The fundamental theorem of calculus gives

$$D(f) = \int_0^1 \psi'(s)\, ds - \frac{\psi'(0) + \psi'(1)}{2},$$

the error of the trapezoidal rule along the diagonal path.

A.2 Proof of Parts (i) to (iv)

Proof. Part (i) is immediate from Equation (A.1), since a game in $\mathcal{G}_{\leq 2}$ has no dividend of degree three or more.

For (ii), one direction follows from (i) together with the linearity of $\mathcal{E}$ and D in the game. For the other, set $z = v - v'$, so that $D(z) = 0$ and it suffices to produce $b \in \mathcal{G}_{\leq 2}$ with $\mathcal{E}(b) = \mathcal{E}(z)$. Write $c_0 = z(\varnothing)$ and let a and $a + d$ be the empty- and full-mask gradients of z. Set

$$b(y) = c_0 + \sum_i a_i y_i + \sum_{i<j} c_{ij} y_i y_j,$$

$$c_{ij} = \frac{d_i + d_j}{g-2} - \frac{\sum_\ell d_\ell}{(g-1)(g-2)},$$

which is well defined because $g \geq 3$. Evaluating at the empty mask gives $b(0) = c_0$, and $\partial_i b(y) = a_i + \sum_{j \neq i} c_{ij} y_j$ gives $\nabla b(0) = a$, so the empty-mask data match.

At the full mask $\partial_i b(\mathbf{1}) = a_i + \sum_{j \neq i} c_{ij}$. Summing the $g-1$ terms of the first fraction and the $g-1$ copies of the second,

$$\begin{aligned} \sum_{j \neq i} c_{ij} &= \frac{(g-1)d_i + \left(\sum_\ell d_\ell - d_i\right)}{g-2} - \frac{\sum_\ell d_\ell}{g-2} \\ &= \frac{(g-2)d_i}{g-2} = d_i, \end{aligned} \tag{A.2}$$

so the full-mask gradients match as well. Summing Equation (A.2) over i counts each unordered pair twice, whence $\sum_{i<j} c_{ij} = \frac{1}{2}\sum_\ell d_\ell$ and

$$b(\mathbf{1}) = c_0 + \sum_i a_i + \tfrac{1}{2}\sum_\ell d_\ell.$$

On the other side, substituting $K_{0i} = a_i$ and $K_{1i} = a_i + d_i$ into the definition of D gives

$$D(z) = z(N) - c_0 - \sum_i a_i - \tfrac{1}{2}\sum_\ell d_\ell,$$

so the hypothesis $D(z) = 0$ says exactly that $z(N) = b(\mathbf{1})$. All four components of $\mathcal{E}(z)$ are therefore realized by b.

Part (ii) identifies the level sets on which a quadratic-invariant rule must be constant, which is (iii). Such a Λ takes the same value at $\mathcal{E}(v)$ and $\mathcal{E}(v+b)$ for every $b \in \mathcal{G}_{\leq 2}$, hence is constant on each coset of $\mathcal{E}(\mathcal{G}_{\leq 2})$ inside the image of $\mathcal{E}$, and by (ii) those cosets are precisely the level sets of D there.

For (iv), a single triple monomial has one dividend, of degree three, so Equation (A.1) gives $D = 1 - \frac{3}{2} = -\frac{1}{2}$. The games $y_1y_2y_3$ and $y_1y_2y_4$ thus share the value of D, while $\hom_3$ sends them to distinct standard basis vectors and hence to distinct lines. The second game requires $g \geq 4$. By (iii) a quadratic-invariant rule returns the same output for both, so it cannot return $\operatorname{span}\{\hom_3(\cdot)\}$ for each. □

The construction in (ii) never uses that the tuple it starts from comes from a game, which yields the following.

Corollary A.1. *For $g \geq 3$ the map $\mathcal{E}$ is onto $\mathbb{R}^{2+2g}$ and $\ker D = \mathcal{E}(\mathcal{G}_{\leq 2})$. Consequently every linear functional of first-order endpoint data that vanishes on $\mathcal{G}_{\leq 2}$ is a scalar multiple of D.*

Proof. Let $x \in \mathbb{R}^{2+2g}$. By Equation (A.1) the game $-2D(x)\, y_1y_2y_3$ has D-value $D(x)$, so

$$x - \mathcal{E}\big(-2D(x)\, y_1y_2y_3\big)$$

is annihilated by D, and the construction in (ii) realizes it as $\mathcal{E}(b)$ for some $b \in \mathcal{G}_{\leq 2}$. Hence x lies in the image of $\mathcal{E}$. Taking $D(x) = 0$ gives $\ker D \subseteq \mathcal{E}(\mathcal{G}_{\leq 2})$, and the reverse inclusion is part (i). Since D is not identically zero, $\ker D$ has codimension one, so every linear functional vanishing on it is proportional to D. □

A.3 Proof of Part (v)

Part (v) combines the factorization through D with the invariance of D under the action of $\mathfrak{S}_g$.

Lemma A.1. *$D(\pi \cdot v) = D(v)$ for every game v and every $\pi \in \mathfrak{S}_g$.*

Proof. Reindexing the dividend sum by $C \mapsto \pi^{-1}(C)$ gives $\mu_{\pi\cdot v}(A) = \mu_v(\pi^{-1}(A))$. Write $\widetilde{v}$ for the multilinear extension of v and P_π for the permutation matrix with $(P_\pi y)_j = y_{\pi(j)}$. Substituting the dividend identity into the expansion,

$$\widetilde{v}(P_\pi y) = \sum_C \mu_v(C) \prod_{j \in C} y_{\pi(j)} = \sum_A \mu_v(\pi^{-1}(A)) \prod_{k \in A} y_k,$$

so the multilinear extension of $\pi \cdot v$ is $\widetilde{v} \circ P_\pi$.

Both 0 and $\mathbf{1}$ are fixed by P_π, so the endpoint values are unchanged, and the chain rule gives $\nabla(\widetilde{v} \circ P_\pi)(0) = P_\pi^\top \nabla\widetilde{v}(0)$ together with the same identity at $\mathbf{1}$. Permuting a vector leaves its coordinate sum alone, so $\sum_i K_{0i}$ and $\sum_i K_{1i}$ are unchanged and so is every term of D. □

Proposition A.1. *Let $g \geq 3$ and let $L \subseteq \mathbb{R}^{\binom{g}{3}}$ be a one-dimensional subspace with $\pi \cdot L = L$ for every $\pi \in \mathfrak{S}_g$. Then $L = \operatorname{span}\{\hom_3(e_3)\}$.*

Proof. The coordinate rule is a left action, since $(\pi \cdot (\sigma \cdot c))_T = c_{\sigma^{-1}\pi^{-1}(T)} = ((\pi\sigma) \cdot c)_T$.

Fix $c \in L$ with $c \neq 0$. For each π the vector $\pi \cdot c$ lies in $L = \mathbb{R}c$, so $\pi \cdot c = \chi(\pi)\, c$ for a scalar $\chi(\pi)$, nonzero because the action is invertible. Applying $(\pi\sigma) \cdot c = \pi \cdot (\sigma \cdot c)$ gives $\chi(\pi\sigma) = \chi(\pi)\chi(\sigma)$, so χ is a homomorphism into $\mathbb{R}^\times$, and every π has finite order, so $\chi(\pi) \in \{\pm 1\}$. Transpositions generate $\mathfrak{S}_g$ and are mutually conjugate, so χ takes one value on all of them and is determined by it. Thus χ is either trivial or the sign character.

Suppose χ were the sign character. Take $T = \{1, 2, 3\}$, which exists because $g \geq 3$, and the transposition $\pi = (1\ 2)$. Since $\pi^{-1}(T) = T$ we get $(\pi \cdot c)_T = c_T$, while $\pi \cdot c = -c$ gives $(\pi \cdot c)_T = -c_T$. Hence $c_T = 0$. For any other triple T' pick σ with $\sigma(T) = T'$, which is possible because $\mathfrak{S}_g$ is transitive on three-element subsets. Then

$$\chi(\sigma)\, c_{T'} = (\sigma \cdot c)_{T'} = c_{\sigma^{-1}(T')} = c_T = 0,$$

and $\chi(\sigma) \neq 0$ forces $c_{T'} = 0$, so $c = 0$, a contradiction.

Therefore χ is trivial, that is $c_{\pi^{-1}(T)} = c_T$ for every π and every triple T. Transitivity again makes all coordinates of c equal, so c is a nonzero multiple of $\hom_3(e_3)$. □

Let Λ be quadratic-invariant and equivariant and let $L = \Lambda(\mathcal{E}(v))$ be a nonzero line of cubic coefficients. Part (iii) factors Λ through D, and Lemma A.1 gives $D(\pi \cdot v) = D(v)$, so $\Lambda(\mathcal{E}(\pi \cdot v)) = \Lambda(\mathcal{E}(v)) = L$. Equivariance rewrites the left side as $\pi \cdot L$, so $\pi \cdot L = L$ for every π, and Proposition A.1 gives part (v).

A.4 Sharpness of the Hypothesis $g \geq 4$

At $g = 3$ there is a single three-element subset, so $\mathbb{R}^{\binom{3}{3}} = \mathbb{R}$ and every nonzero cubic coefficient vector spans the same line. The constant rule $\Lambda \equiv \operatorname{span}\{\hom_3(e_3)\}$ is quadratic-invariant because it is constant, equivariant because the permutation action on a lone triple is trivial, and correct on every game with $\hom_3(v) \neq 0$. Part (iv) therefore fails at $g = 3$ and holds for every $g \geq 4$, and the counterexample shows where the fourth player is spent. The games $y_1y_2y_3$ and $y_1y_2y_4$ are indistinguishable to D and span different lines.

B Cubic Hermite Interpolation

The Hermite identity of the main paper rests on a cubic polynomial that reproduces the endpoint data. This appendix constructs it, computes the Shapley value of its Boolean restriction, and proves that distinct weights select distinct cubic lines once $g \geq 4$.

Throughout, w has strictly positive coordinates summing to one, and $\mathcal{E}(f) = (v_0, v_1, K_0, K_1)$ are the first-order endpoint data of a differentiable mask extension f. Only that extension occurs below, so we write $D = D(f)$. Put

$$\xi(y) = w^\top y, \qquad \Delta_i = K_{1i} - K_{0i}, \qquad S_\Delta = \sum_i \Delta_i,$$

so that $\frac{1}{2}\sum_i (K_{0i} + K_{1i}) = \sum_i K_{0i} + \frac{1}{2} S_\Delta$ and

$$D = v_1 - v_0 - \sum_i K_{0i} - \tfrac{1}{2} S_\Delta. \tag{B.1}$$

Since w lies in the simplex, ξ maps the cube into $[0, 1]$, with the empty and full masks at the two ends.

The game $\prod_{i \in A} y_i$ assigns $1/|A|$ to each $i \in A$ and nothing to the remaining groups, because i is pivotal exactly when it comes last among A in a uniformly random order. By linearity the Shapley value of a game splits each dividend equally among the members of the set carrying it. For a polynomial a on $[0, 1]^g$ we write $\phi(a)$ for the Shapley value of the game $S \mapsto a(\mathbf{1}_S)$.

B.1 THE INTERPOLANT

The interpolant adds a blending term to a polynomial of degree two, giving

$$h(y) = v_0 + \sum_i K_{0i} y_i + \xi(y) \sum_i \Delta_i y_i \\ - \tfrac{1}{2}\xi(y)^2 S_\Delta + \big(3\xi(y)^2 - 2\xi(y)^3\big) D. \tag{B.2}$$

Proposition B.1. $h(0) = v_0$, $h(\mathbf{1}) = v_1$, $\nabla h(0) = K_0$, *and* $\nabla h(\mathbf{1}) = K_1$.

Proof. The terms of Equation (B.2) before the last match both endpoint gradients and leave a discrepancy D at the full mask. The blending function $3\xi^2 - 2\xi^3$ removes that discrepancy without disturbing the gradients. It equals 0 at $\xi = 0$ and 1 at $\xi = 1$, while its derivative $6\xi(1-\xi)$ vanishes at both.

The empty-mask value is $h(0) = v_0$, since every other term of Equation (B.2) vanishes at $y = 0$. Differentiating,

$$\partial_i h(y) = K_{0i} + w_i \sum_j \Delta_j y_j + \xi(y)\Delta_i \\ - \xi(y) w_i S_\Delta + 6\xi(y)\big(1 - \xi(y)\big) w_i D, \tag{B.3}$$

and every term after K_{0i} carries a factor $\xi(y)$ or $\sum_j \Delta_j y_j$, both of which vanish at $y = 0$. Hence $\nabla h(0) = K_0$.

At $y = \mathbf{1}$ we have $\xi = 1$, so

$$h(\mathbf{1}) = v_0 + \sum_i K_{0i} + S_\Delta - \tfrac{1}{2} S_\Delta + D = v_1$$

by Equation (B.1). Setting $y = \mathbf{1}$ in Equation (B.3), the terms $w_i S_\Delta$ and $-w_i S_\Delta$ cancel and the blending derivative vanishes, leaving $\partial_i h(\mathbf{1}) = K_{0i} + \Delta_i = K_{1i}$. □

B.2 THE BOOLEAN RESTRICTION

Let $\mathrm{ml}(h)$ be the unique multilinear polynomial agreeing with h on $\{0,1\}^g$, that is, the multilinear extension of the game obtained by restricting h to the cube. Reducing y_i^2 to y_i gives

$$\xi \sum_j \Delta_j y_j = \sum_i w_i \Delta_i y_i \\ + \sum_{i<j} (w_i \Delta_j + w_j \Delta_i) y_i y_j, \tag{B.4}$$

$$\xi^2 = \sum_i w_i^2 y_i + 2 \sum_{i<j} w_i w_j y_i y_j, \tag{B.5}$$

$$\xi^3 = \sum_i w_i^3 y_i + 3 \sum_{i<j} w_i w_j (w_i + w_j) y_i y_j \\ + 6 \sum_{i<j<k} w_i w_j w_k y_i y_j y_k. \tag{B.6}$$

A monomial with a repeated index loses degree under the reduction, so the 3! orderings of i, j, k in the expansion of ξ^3 are the only source of $y_i y_j y_k$. Only the last term of Equation (B.2) therefore reaches degree three, contributing $-2D \cdot 6 w_i w_j w_k$ to the dividend of $\{i, j, k\}$, so that

$$\hom_3(\mathrm{ml}(h)) = -12\, D\ \hom_3(p_w).$$

The same multilinearizations give the attribution of the interpolant itself.

Proposition B.2. *The Boolean restriction of h has Shapley value*

$$\phi_i(h) = \tfrac{1}{2}(K_{0i} + K_{1i}) + w_i\big(1 + w_i - \|w\|_2^2\big) D.$$

Proof. Apply the dividend rule to Equation (B.2) term by term. The constant v_0 contributes nothing, and $\sum_j K_{0j} y_j$ contributes K_{0i}. From Equation (B.4),

$$\begin{aligned}\phi_i\Big(\xi \sum_j \Delta_j y_j\Big) &= w_i \Delta_i + \tfrac{1}{2} \sum_{j \neq i} (w_i \Delta_j + w_j \Delta_i) \\ &= \tfrac{1}{2} \Delta_i + \tfrac{1}{2} w_i S_\Delta,\end{aligned}$$

using $\sum_{j \neq i} \Delta_j = S_\Delta - \Delta_i$ and $\sum_{j \neq i} w_j = 1 - w_i$. Equation (B.5) gives $\phi_i(\xi^2) = w_i^2 + w_i(1 - w_i) = w_i$, so the term $-\frac{1}{2}\xi^2 S_\Delta$ contributes $-\frac{1}{2} w_i S_\Delta$, canceling the $\frac{1}{2} w_i S_\Delta$ just obtained. For the cubic, Equation (B.6) gives

$$\begin{aligned}\phi_i(\xi^3) = w_i^3 &+ \tfrac{3}{2} \sum_{j \neq i} w_i w_j (w_i + w_j) \\ &+ 2 w_i \sum_{\substack{j<k \\ j,k \neq i}} w_j w_k \\ &= w_i \Big(1 + \tfrac{\|w\|_2^2 - w_i}{2}\Big),\end{aligned}$$

where the last step uses $\sum_{j \neq i} w_j = 1 - w_i$, $\sum_{j \neq i} w_j^2 = \|w\|_2^2 - w_i^2$, and $2 \sum_{j<k,\, j,k \neq i} w_j w_k = (1 - w_i)^2 - \big(\|w\|_2^2 - w_i^2\big)$. The blending term therefore contributes

$$\big(3\phi_i(\xi^2) - 2\phi_i(\xi^3)\big) D = w_i \big(1 + w_i - \|w\|_2^2\big) D,$$

and adding $K_{0i} + \frac{1}{2}\Delta_i = \frac{1}{2}(K_{0i} + K_{1i})$ gives the formula. □

Since $\sum_i w_i(1 + w_i - \|w\|_2^2) = 1 + \|w\|_2^2 - \|w\|_2^2 = 1$, the coordinates of $\phi(h)$ sum to $\sum_i \frac{1}{2}(K_{0i} + K_{1i}) + D = v_1 - v_0$, as efficiency requires. Under uniform weights the correction is D/g, so the attribution is the endpoint-gradient average plus an equal share of D.

B.3 Distinct Weights, Distinct Lines

Proposition B.3. *Let $g \geq 4$, and let w and $\widetilde{w}$ have strictly positive coordinates summing to one. If $\operatorname{span}\{p_w\} = \operatorname{span}\{p_{\widetilde{w}}\}$ then $w = \widetilde{w}$.*

Proof. Equality of the spans makes the strictly positive vectors $\hom_3(p_w)$ and $\hom_3(p_{\widetilde{w}})$ proportional. Given two indices $i \neq i'$, choose distinct j, k outside $\{i, i'\}$, which $g \geq 4$ makes possible. Taking the ratio of the entries at $\{i, j, k\}$ and at $\{i', j, k\}$ cancels the constant of proportionality together with $w_j w_k$ and $\widetilde{w}_j \widetilde{w}_k$, so $w_i \widetilde{w}_{i'} = w_{i'} \widetilde{w}_i$. All the ratios $w_i / \widetilde{w}_i$ therefore agree, and since both vectors sum to one their common value is one. □

At $g = 3$ there is a single triple and every p_w spans the same line, so $g \geq 4$ cannot be weakened.

C Constrained Partial Regression and Generalized Ridge

The main paper describes the estimator through objects in the sampled response space. This appendix gives the coefficient-level construction behind them, the numerical rules the algorithm applies, and the proof of its guarantee theorem.

C.1 The Constrained Fit

The feasible set of Equation (8) is an affine subspace. The minimum-norm point of $\{a : \mathbf{1}^\top a = d\}$ is $a_k^0(d) = d\,\mathbf{1} / \|\mathbf{1}\|_2^2$, and every feasible a has the form $a_k^0(d_z) + P_k \theta$. Substituting and writing

$$\widetilde{z}_k = W\big(\bar{z}_\Omega - X_k a_k^0(d_z)\big) \tag{C.1}$$

turns the problem into $\min_\theta \|A_k\theta - \widetilde{z}_k\|_2^2$, whose minimum-norm solution is $A_k^\dagger \widetilde{z}_k$. That vector lies in $\mathrm{range}(A_k^\top) \subseteq \mathrm{range}(P_k)$ and so is fixed by P_k, giving

$$C_k z = a_k^0(d_z) + P_k A_k^\dagger \widetilde{z}_k. \tag{C.2}$$

The two summands are orthogonal, so Equation (C.2) is the feasible minimizer of least norm. Both d_z and $\bar{z}_\Omega$ are linear in z, so C_k is linear. With $H_2 = A_2 A_2^\dagger$ and $W X_2 P_2 = A_2$,

$$\rho_2(z) = W\big(\bar{z}_\Omega - X_2 C_2 z\big) = (I - H_2)\widetilde{z}_2. \tag{C.3}$$

In particular $H_2 r = 0$. The residual is thus taken from $\widetilde{z}_2$ rather than from $W\bar{z}_\Omega$. The affine constraint displaces the point being projected by $W X_2 a_2^0(d_z)$, which is proportional to d_z.

The estimability condition buys exactness on $\mathcal{G}_{\leq 2}$.

Proposition C.1. *If* $\ker(A_2) \subseteq \ker(\Phi_2 P_2)$*, then* $\mathcal{A}_2 b = \phi(b)$ *for every* $b \in \mathcal{G}_{\leq 2}$.

Proof. Compare the fitted coefficient vector with the exact one. Let c_b be the coefficient vector of $\bar{b}$, so that $X_2 c_b = \bar{b}_\Omega$, $\mathbf{1}^\top c_b = d_b$ and $\phi(b) = \Phi_2 c_b$. It is feasible and has zero residual, so the minimum in Equation (8) is zero and $C_2 b$ has zero residual as well. Their difference $\theta = C_2 b - c_b$ therefore satisfies $\mathbf{1}^\top \theta = 0$ and $W X_2 \theta = 0$, so $P_2\theta = \theta$ and $A_2\theta = 0$, whence $\mathcal{A}_2 b - \phi(b) = \Phi_2\theta = \Phi_2 P_2 \theta = 0$. □

The same condition at degree three, $\ker(A_3) \subseteq \ker(\Phi_3 P_3)$, would let the sampled values determine the Shapley value of every game in $\mathcal{G}_{\leq 3}$. It fails at every confirmatory budget. For $g = 8$ and the designs the pinned PolySHAP sampler draws with the settings and random states of our experiments, 50 per budget, it holds in none at any $B \leq 112$, which includes the confirmatory budgets 48, 64 and 80, in 25 at $B = 128$ and in all 50 at $B = 144$ and $B = 160$, whereas the degree-two condition holds in every one of these designs at the confirmatory budgets, although $\mathrm{rank}(A_2) \leq 30$ at $B = 48$ (Appendix E). With p complementary pairs the rank of A_3 is at most $\min(p, 28) + \min(p, 63)$, below the 91 free directions until $p = 63$. Structured designs behave differently. The coalitions of sizes 0, 1 and 2 together with their complements determine the Shapley value of every game of interaction order at most four (Hu & Wang, 2023).

C.2 Proof of Theorem 2

For fixed β, the objective and the constraint of Equation (10) are those of Equation (8) for $v - \beta q$, since $d_{v-\beta q} = d_v - \beta d_q$. The inner minimum is therefore attained at $C_2(v - \beta q)$, whose residual is $\rho_2(v - \beta q) = r - \beta u$ by linearity, and minimizing the profiled objective $\|r - \beta u\|_2^2 + \tau\beta^2$ gives $\widehat{\beta}_\tau = u^\top r/(\delta + \tau)$. Reading the attribution off the fitted model gives $\Phi_2 C_2(v - \widehat{\beta}_\tau q) + \widehat{\beta}_\tau \phi(q) = \mathcal{A}_2 v + \widehat{\beta}_\tau \Delta_q$ by linearity of $\mathcal{A}_2$. Both $\phi(q)$ and $\mathcal{A}_2 q$ sum to d_q, so Δ_q sums to zero and the attribution is efficient for every τ. Exactness at $\tau = 0$ is the case $e = 0$ of Proposition C.2 below, together with Proposition C.1, which turns the quadratic fit into $\phi(b)$.

C.3 Partial Regression Away from the Exact Case

When $\delta > 0$ and the coefficient of q is left unpenalized, the main paper obtains $\widehat{\beta}_0 = u^\top r/\delta$ and $\widehat{\phi}_q(v) = \mathcal{A}_2 v + \widehat{\beta}_0 \Delta_q$. The next statement records what a departure from $\mathcal{G}_{\leq 2} \oplus \mathrm{span}\{q\}$ costs. At $e = 0$ it gives $\widehat{\beta}_0 = \beta_\star$ and $\widehat{\phi}_q(v) = \phi(v)$, the exactness asserted by Theorem 2 of the main paper.

Proposition C.2. *Assume* $\ker(A_2) \subseteq \ker(\Phi_2 P_2)$ *and* $\delta > 0$*, and let* $v = b + \beta_\star q + e$ *with* $b \in \mathcal{G}_{\leq 2}$*. Then*

$$\widehat{\beta}_0 - \beta_\star = \frac{u^\top \rho_2(e)}{\delta},$$

$$\widehat{\phi}_q(v) - \phi(v) = \mathcal{A}_2 e - \phi(e) + (\widehat{\beta}_0 - \beta_\star)\Delta_q,$$

and consequently

$$\left\|\widehat{\phi}_q(v) - \phi(v)\right\|_2 \leq \|\mathcal{A}_2 e - \phi(e)\|_2 + \frac{\|\rho_2(e)\|_2}{\sqrt{\delta}}\ \|\Delta_q\|_2 .$$

Proof. Linearity of ρ_2 and $\rho_2(b) = 0$ give $r = \beta_\star u + \rho_2(e)$, and taking the inner product with u gives the first identity. For the second, linearity of $\mathcal{A}_2$ and ϕ gives

$$\widehat{\phi}_q(v) - \phi(v) = \big(\mathcal{A}_2 b - \phi(b)\big) + \big(\widehat{\beta}_0 - \beta_\star\big)\Delta_q + \mathcal{A}_2 e - \phi(e),$$

and the first bracket vanishes by Proposition C.1. The bound follows from $\big|u^\top \rho_2(e)\big| \le \|u\|_2 \, \|\rho_2(e)\|_2$ and $\|u\|_2 = \sqrt{\delta}$. □

Proposition C.3. *Let $\delta > 0$ and let the coefficient of q be unpenalized. Replacing q by αq for any $\alpha \neq 0$ leaves $\widehat{\phi}_q(v)$ unchanged.*

Proof. By linearity of ρ_2, $\mathcal{A}_2$ and ϕ, the replacement sends u to αu, δ to $\alpha^2\delta$ and Δ_q to $\alpha\Delta_q$, hence $\widehat{\beta}_0$ to $\widehat{\beta}_0/\alpha$. The product $\widehat{\beta}_0\Delta_q$ is therefore unchanged, and so is $\mathcal{A}_2 v + \widehat{\beta}_0\Delta_q$. □

C.4 The Residualized Cubic Space

Let E embed degree-two coefficient vectors into degree-three ones by padding the triple coordinates with zeros, so that

$$X_3 E = X_2, \qquad \mathbf{1}^\top E = \mathbf{1}^\top, \qquad \Phi_3 E = \Phi_2. \tag{C.4}$$

The map $\Pi_{2\leftarrow 3}$ sends a degree-three coefficient direction to its efficiency-constrained quadratic refit, and the vector η_q is the residualized coefficient vector of the supplied regressor,

$$\Pi_{2\leftarrow 3} = E P_2 A_2^\dagger A_3, \qquad \eta_q = c_q - E C_2 q, \tag{C.5}$$

where c_q is the exact coefficient vector of $\bar{q}$. Both terms of η_q have full-coalition value d_q, so $\mathbf{1}^\top \eta_q = 0$ and hence $P_3\eta_q = \eta_q$. Therefore, using Equation (C.4),

$$A_3\eta_q = W\big(\bar{q}_\Omega - X_2 C_2 q\big) = u, \qquad \Phi_3\eta_q = \Delta_q. \tag{C.6}$$

For $\delta > 0$ put $P_u = uu^\top/\delta$ and define the coefficient residual map R_q, the residualized cubic design B_q and the attribution map $\mathcal{D}_q$ by

$$\begin{aligned} R_q &= P_3 - \Pi_{2\leftarrow 3} - \eta_q \frac{u^\top A_3}{\delta}, \\ B_q &= A_3 R_q, \qquad \mathcal{D}_q = \Phi_3 R_q. \end{aligned} \tag{C.7}$$

Proposition C.4. *$A_3 R_q = (I - H_2 - P_u)A_3$, every column of $\mathcal{D}_q$ has coordinate sum zero, and $\mathrm{range}(A_2)$, $\mathrm{span}\{u\}$ and $\mathrm{range}(B_q)$ are mutually orthogonal.*

Proof. Since $\mathbf{1}^\top E P_2 = \mathbf{1}^\top P_2 = 0$, the columns of EP_2 lie in $\mathbf{1}^\perp$, so $P_3 E P_2 = EP_2$ and $A_3 E P_2 = W X_3 E P_2 = W X_2 P_2 = A_2$. Hence $A_3\Pi_{2\leftarrow 3} = A_2 A_2^\dagger A_3 = H_2 A_3$. With $A_3 P_3 = A_3$ and Equation (C.6),

$$A_3 R_q = A_3 - H_2 A_3 - u\frac{u^\top A_3}{\delta} = (I - H_2 - P_u)A_3.$$

All three terms of R_q map into $\mathbf{1}^\perp$, so $\mathbf{1}^\top R_q = 0$, and $\mathbf{1}^\top \Phi_3 = \mathbf{1}^\top$ gives $\mathbf{1}^\top \mathcal{D}_q = 0$. Finally Equation (C.3) makes u orthogonal to $\mathrm{range}(A_2)$, so $H_2 + P_u$ is the orthogonal projector onto $\mathrm{range}(A_2) \oplus \mathrm{span}\{u\}$ and the range of $(I - H_2 - P_u)A_3$ is orthogonal to both. □

When the retention test of Equation (C.15) fails, the algorithm drops the last term of R_q, leaving

$$\begin{aligned} R_0 &= P_3 - \Pi_{2\leftarrow 3}, \qquad \mathcal{D}_0 = \Phi_3 R_0, \\ B_0 &= A_3 R_0 = (I - H_2)A_3, \end{aligned} \tag{C.8}$$

where the second line and $\mathbf{1}^\top \mathcal{D}_0 = 0$ come from the same computation. This branch never forms P_u, B_q or $\mathcal{D}_q$, so it does not depend on q.

C.5 Candidates, Selection and Shrinkage

Let $B_q = U \operatorname{diag}(s_j) V^\top$ be a compact singular value decomposition with singular values $s_1 \geq \cdots \geq s_{\operatorname{rank}(B_q)} > 0$. For $\lambda \in [0, \infty]$ the ridge solution of $\min_\theta \|B_q\theta - r\|_2^2 + \lambda \|\theta\|_2^2$ is

$$\theta_\lambda = V \operatorname{diag}\Big(\frac{s_j}{s_j^2 + \lambda}\Big) U^\top r, \tag{C.9}$$

with the filter read as zero at $\lambda = \infty$. The ridge correction is then $\mathcal{D}_q \theta_\lambda$, and at rank zero the factors have no columns and that correction vanishes for every λ.

A candidate pairs the ridge parameter λ with a shrinkage $\gamma \in [0, 1]$ on the coefficient of the supplied regressor. Because $H_2 r = 0$, the fitted part of r at a candidate (γ, λ) is $S_{\gamma,\lambda} r$ for the fixed smoother

$$S_{\gamma,\lambda} = H_2 + \gamma P_u + U \operatorname{diag}\Big(\frac{s_j^2}{s_j^2 + \lambda}\Big) U^\top, \tag{C.10}$$

whose trace is the effective degrees of freedom,

$$\mathrm{df}_{\gamma,\lambda} = \operatorname{rank}(A_2) + \gamma + \sum_j \frac{s_j^2}{s_j^2 + \lambda}. \tag{C.11}$$

A candidate is *admissible* when $m - \mathrm{df}_{\gamma,\lambda} > 0$, and admissible candidates are scored by

$$\begin{aligned} \mathrm{GCV}(\gamma, \lambda) &= \frac{\mathrm{RSS}_{\gamma,\lambda}}{(m - \mathrm{df}_{\gamma,\lambda})^2}, \\ \mathrm{RSS}_{\gamma,\lambda} &= \|(I - S_{\gamma,\lambda}) r\|_2^2 . \end{aligned} \tag{C.12}$$

Multiplying Equation (C.12) by m gives the other common normalization and selects the same candidate.

The grid is fixed in advance, with

$$\gamma \in \{0, .05, .10, .20, .35, .50, .65, .80, .90, .95, 1\}$$

and $\lambda / s_1^2 \in \{\infty, 0\} \cup \{10^{-8}, 10^{-7.5}, \ldots, 10^{8}\}$, giving $11 \times 35 = 385$ pairs. The normalized ridge grid needs s_1 and is therefore used only at positive rank. At rank zero every ridge label carries the same zero correction. The implementation traverses γ in the displayed order and, for each γ, traverses λ in its displayed order, retaining the incumbent candidate whenever the two scores differ by at most 64 times the machine epsilon times the maximum of the two score magnitudes and tiny, the smallest positive normal double-precision number. On the omitted branch the same rules apply with B_0 and $\mathcal{D}_0$ in place of B_q and $\mathcal{D}_q$, and with γP_u dropped from Equation (C.10) and γ from Equation (C.11). The cubic-ridge control of the main paper runs that branch on every target instead of reaching it by the retention test.

The decompositions above are shared across all targets of one mask design, so beyond the coalition values and the two endpoint solves each further target costs one rank-one update and a scan of the 385 candidates.

Write $c_{\gamma,\lambda} = \gamma \widehat{\beta}_0 \Delta_q + \mathcal{D}_q \theta_\lambda$ for the nonquadratic correction of a candidate. It is linear in the residual, $c_{\gamma,\lambda} = L_{\gamma,\lambda} r$, with

$$L_{\gamma,\lambda} = \gamma \Delta_q \frac{u^\top}{\delta} + \mathcal{D}_q V \operatorname{diag}\Big(\frac{s_j}{s_j^2 + \lambda}\Big) U^\top. \tag{C.13}$$

Setting $\widehat{\sigma}^2 = \mathrm{RSS}_{\gamma,\lambda} / (m - \mathrm{df}_{\gamma,\lambda})$, the output at the selected candidate is

$$\widehat{\phi} = \mathcal{A}_2 v + \omega\, c_{\gamma,\lambda}, \qquad \omega = \Big(1 - \frac{\widehat{\sigma}^2 \|L_{\gamma,\lambda}\|_F^2}{\|c_{\gamma,\lambda}\|_2^2}\Big)_+, \tag{C.14}$$

with $\omega = 0$ when the correction vanishes. The factor is a numerical regularizer. The map $L_{\gamma,\lambda}$ belongs to one fixed candidate, but selection makes the estimator as a whole nonlinear in the response.

Algorithm 1 Sensitivity-informed cubic Shapley regression. Neither the exact $\phi(v)$ nor any coalition value beyond the B inputs enters the construction, the selection or the output. The Shapley vector of the supplied basis function is available in closed form (Section 3 of the main paper).

Input: sampled interior coalitions Ω with values v_Ω, the endpoint values $v(\varnothing)$, $v(N)$ and gradients K_0, K_1, and the sampling normalization W
Parameter: ε, the candidate grid and the retention threshold, all fixed in advance
Output: an efficient Shapley estimate $\widehat{\phi}$

1: Form κ, w, p_w and q from K_0 and K_1
2: Fit the efficiency-constrained quadratic model to obtain $\mathcal{A}_2 v$
3: Compute r, u, δ and Δ_q
4: **if** u passes the retention test Equation (C.15) **then** set $\widehat{\beta}_0 = u^\top r/\delta$ **else** omit the supplied-regressor block
5: Build B_q and $\mathcal{D}_q$ on the retained branch, or B_0 and $\mathcal{D}_0$ on the omitted one
6: Score the admissible candidates of the grid by Equation (C.12)
7: Take the minimum-scoring candidate, apply the positive-part factor of Equation (C.14) to its nonquadratic correction, and return the result

The retention test of the main paper is

$$\frac{\delta}{\max\{\|W\bar{q}_\Omega\|_2^2,\ \mathrm{tiny}\}} > 10^{-12}, \tag{C.15}$$

where tiny guards a vanishing denominator. Above that guard the statistic is $\delta/\|W\bar{q}_\Omega\|_2^2$, which $q \mapsto \alpha q$ leaves unchanged as long as the rescaled denominator stays above it too. The test fails whenever $\delta = 0$, which is exactly when q is aliased with the sampled quadratic model after efficiency is enforced. The results below concern exact arithmetic, where the enumeration order settles ties. The tolerance stated with the grid absorbs floating-point noise in the implementation.

Proposition C.5. *For a fixed candidate with $\delta > 0$ and $0 < \lambda < \infty$,*

$$\|L_{\gamma,\lambda}\|_2 \le \gamma\frac{\|\Delta_q\|_2}{\sqrt{\delta}} + \frac{\|\mathcal{D}_q V\|_2}{2\sqrt{\lambda}}.$$

Proof. The first term of Equation (C.13) is a rank-one matrix of spectral norm $\gamma\|\Delta_q\|_2\|u\|_2/\delta = \gamma\|\Delta_q\|_2/\sqrt{\delta}$. In the second, U has orthonormal columns and $\sup_{s\ge 0} s/(s^2+\lambda) = 1/(2\sqrt{\lambda})$. □

C.6 The Algorithm

The main paper describes the estimator in prose. Algorithm 1 collects it, and the proof below refers to it by name.

C.7 Proof of the Guarantee Theorem

Theorem C.1. *Suppose $q \in \mathcal{G}_{\le 3}$, and suppose the branch chosen by the retention test has an admissible candidate in its grid.*

(i) The estimator uses no coalition value beyond the B supplied as input.

(ii) Every admissible candidate, the selected candidate and the scaled output are efficient.

(iii) If the supplied block is retained, $\ker(A_2) \subseteq \ker(\Phi_2 P_2)$, an admissible candidate with $\gamma = 1$ exists, and $v \in \mathcal{G}_{\le 2} \oplus \mathrm{span}\{q\}$, then the selected candidate and the output both equal $\phi(v)$.

(iv) Relabeling the players relabels the output by the same permutation.

Proof. Part (i). Every object Algorithm 1 forms is a function of the supplied values v_Ω, $v(\varnothing)$ and $v(N)$, the fixed design (Ω, W), the endpoint gradients and the closed form $\phi(q)$. Equation (C.12) and Equation (C.14) read only r and the design.

Part (ii). $\mathbf{1}^\top \Delta_q = 0$ because $\phi(q)$ and $\mathcal{A}_2 q$ both sum to d_q, and $\mathbf{1}^\top \mathcal{D}_q = 0$ by Proposition C.4, with $\mathcal{D}_0$ in place of $\mathcal{D}_q$ on the omitted branch. So every $c_{\gamma,\lambda}$ has coordinate sum zero, as does $\omega c_{\gamma,\lambda}$, and $\mathcal{A}_2 v$ is efficient. Choosing among candidates does not disturb this.

Part (iii). Retention gives $\delta > 0$. Write $v = b + \beta_\star q$ with $b \in \mathcal{G}_{\le 2}$. Then $\rho_2(b) = 0$, so $r = \beta_\star u$ and $\widehat{\beta}_0 = \beta_\star$. Since $(I - H_2 - P_u)u = 0$ and $I - H_2 - P_u$ is symmetric, $B_q^\top r = \beta_\star A_3^\top (I - H_2 - P_u)u = 0$, so $U^\top r = 0$ and $\theta_\lambda = 0$ for every λ. The residual of a candidate is therefore $(1-\gamma)\beta_\star u$ and $\mathrm{RSS}_{\gamma,\lambda} = (1-\gamma)^2 \beta_\star^2 \delta$.

If $\beta_\star \neq 0$, an admissible candidate with $\gamma = 1$ has zero score while every admissible $\gamma < 1$ has positive score, so the selection has $\gamma = 1$. Its residual sum of squares is zero, hence $\widehat{\sigma}^2 = 0$ and $\omega = 1$ unless $c_{\gamma,\lambda} = 0$. Either way the output is $\mathcal{A}_2 v + \beta_\star \Delta_q$. Expanding it,

$$\begin{aligned} \mathcal{A}_2 v + \beta_\star \Delta_q &= \mathcal{A}_2 b + \beta_\star \mathcal{A}_2 q + \beta_\star \big(\phi(q) - \mathcal{A}_2 q\big) \\ &= \mathcal{A}_2 b + \beta_\star \phi(q), \end{aligned}$$

and Proposition C.1 turns $\mathcal{A}_2 b$ into $\phi(b)$, so the output is $\phi(v)$. If $\beta_\star = 0$ then $r = 0$, every correction vanishes, and the output is $\mathcal{A}_2 b = \phi(b) = \phi(v)$.

Part (iv). A relabeling of the players induces a permutation of the monomial coordinates and, through it, permutes the columns of every X_k and A_k. The sampled values and W are untouched, since relabeling the masks at the same time leaves each row's value and coalition size alone. Everything above is assembled from Euclidean norms, pseudoinverses, singular values and a fixed traversal order, all of which commute with such a permutation, so Equation (C.15), the selected candidate and ω are unchanged, and Φ_k carries the permutation to the players of the output. □

Proposition C.6. *Let $\alpha \neq 0$ and suppose that replacing q by αq leaves the retention decision unchanged. Then it also leaves the selected pair (γ, λ), the factor ω and the output unchanged.*

Proof. If both decisions omit the supplied block, the candidates are built from r, B_0 and $\mathcal{D}_0$, none of which involves q, and there is nothing to prove. Suppose instead that both retain it. The replacement sends $u \mapsto \alpha u$, $\delta \mapsto \alpha^2 \delta$, $\eta_q \mapsto \alpha \eta_q$ and $\Delta_q \mapsto \alpha \Delta_q$. In Equation (C.7) the last term becomes $\alpha \eta_q (\alpha u^\top A_3)/(\alpha^2 \delta)$, so P_u, R_q, B_q and $\mathcal{D}_q$ are unchanged, and with them every singular value, smoother, parameter count and score. Finally $\widehat{\beta}_0 \mapsto \widehat{\beta}_0/\alpha$ leaves $\gamma \widehat{\beta}_0 \Delta_q$ and Equation (C.13) unchanged, hence also ω and the output. □

The hypothesis is a floating-point matter rather than a mathematical restriction. Both δ and $\|W \bar{q}_\Omega\|_2^2$ carry the factor α^2, so the two decisions agree whenever the guard tiny in Equation (C.15) is inactive at both scalings.

D Implicit Differentiation

The main paper states that the endpoint gradients of an implicit model cost two linear solves. This appendix records the hypothesis behind that statement, verifies it for the graph-classification models, and gives the readout derivative and the mask parameterization.

D.1 The Equilibrium Derivative

Proposition D.1. *Let $\mathcal{U} \subseteq \mathbb{R}^g$ be open and let $F : \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}^n$ be continuously differentiable with*

$$\|\partial_x F(x,\theta)\|_{\mathrm{op}} \le L < 1 \qquad \text{for all } (x,\theta) \in \mathbb{R}^n \times \mathcal{U}. \tag{D.1}$$

Then $x = F(x,\theta)$ has a unique solution $x^\star(\theta)$ for each $\theta \in \mathcal{U}$, the map $x^\star$ is continuously differentiable, and $Z^\star(\theta) = \partial_\theta x^\star(\theta) \in \mathbb{R}^{n \times g}$ is the unique solution of

$$Z = \partial_x F(x^\star, \theta)\, Z + \partial_\theta F(x^\star, \theta). \tag{D.2}$$

Moreover the iteration $Z \mapsto \partial_x F\, Z + \partial_\theta F$ converges to $Z^\star$ from any starting point, with error contracting by a factor L per step in Frobenius norm.

Proof. By Equation (D.1) and the mean value inequality on the convex set $\mathbb{R}^n$, each $F(\cdot, \theta)$ is L-Lipschitz in x, so Banach's theorem gives a unique fixed point $x^\star(\theta)$.

Write $\Psi(x, \theta) = x - F(x, \theta)$, so $\partial_x \Psi = I - \partial_x F$. The Neumann series $\sum_{k \geq 0} (\partial_x F)^k$ converges in operator norm and inverts $I - \partial_x F$, with $\left\| (I - \partial_x F)^{-1} \right\|_{\mathrm{op}} \leq (1 - L)^{-1}$. The implicit function theorem therefore applies at every $(x^\star(\theta), \theta)$ and produces a C^1 local branch of solutions. Every value of that branch solves $x = F(x, \theta)$, and the solution is unique, so the branch coincides with $x^\star$. Hence $x^\star$ is continuously differentiable. Differentiating $x^\star = F(x^\star, \theta)$ gives

$$\partial_\theta x^\star = (I - \partial_x F)^{-1} \partial_\theta F,$$

which is Equation (D.2) rearranged, and invertibility of $I - \partial_x F$ makes its solution unique.

For the iteration, $\|MZ\|_F \leq \|M\|_{\mathrm{op}} \|Z\|_F$ for any conformable M and Z, so $Z \mapsto \partial_x F \, Z + \partial_\theta F$ is an L-contraction on $(\mathbb{R}^{n \times g}, \|\cdot\|_F)$ and Banach's theorem applies again. □

Global uniqueness is what keeps the argument short, and a bound of the form $\|\partial_x F(x^\star(\theta), \theta)\|_{\mathrm{op}} < 1$ imposed only along a selected branch does not supply it. For $F(x, \theta) = (3x - x^3)/2$ the points ± 1 are fixed with vanishing derivative, so the selection returning -1 for $\theta_1 < 0$ and 1 otherwise meets that bound while being discontinuous. That example violates Equation (D.1), which also covers its third fixed point at 0.

Two features of Equation (D.2) govern the forward-mode cost. The operator $I - \partial_x F$ does not depend on which mask direction is differentiated, so all g columns of $Z^\star$ come from one system with g right-hand sides. And $x^\star(\theta)$ enters only as the point at which the Jacobians are evaluated, so no further equilibrium solve is needed once that point is known.

In what follows $\mathcal{U}$ is an open neighborhood of $[0, 1]^g$, so that $x^\star(0)$ and $x^\star(\mathbf{1})$ are both covered.

D.2 The Readout, and Why f Extends the Game

Let ℓ be a continuously differentiable scalar readout and put $f(\theta) = \ell(x^\star(\theta), \theta)$. The chain rule gives

$$\nabla f(\theta) = Z^\star(\theta)^\top \nabla_x \ell(x^\star(\theta), \theta) + \nabla_\theta \ell(x^\star(\theta), \theta).$$

In both model families the readout is a fixed pooling and linear head applied to the equilibrium state, so it does not depend on θ directly and the second term vanishes. Writing $x_0 = x^\star(0)$, $x_1 = x^\star(\mathbf{1})$, $Z_0 = Z^\star(0)$ and $Z_1 = Z^\star(\mathbf{1})$, the endpoint gradients are

$$K_0 = Z_0^\top \nabla_x \ell(x_0), \qquad K_1 = Z_1^\top \nabla_x \ell(x_1).$$

That f is a differentiable mask extension of v follows from how the mask is applied. In both families a coordinate belonging to group i is replaced by the convex combination of its observed value and its baseline, with weight θ_i on the observed value. At $\theta = \mathbf{1}_S$ this retains exactly the groups in S and sets the others to their baseline, which is the input defining $v(S)$. Hence $f(\mathbf{1}_S) = v(S)$. The mask-to-input map is affine, so $\partial_\theta F$ is obtained by composing the model's input Jacobian with a constant matrix.

D.3 Cost

The empty and full coalitions are two of the B evaluations in the budget, and computing them already produces x_0 and x_1. Obtaining the endpoint gradients therefore adds one linear solve at each of those points, together with the Jacobian and readout applications it entails. No interior coalition is evaluated and no further equilibrium is computed.

Two routes give the same gradient. The forward system Equation (D.2) has g right-hand sides. Because the readout is scalar, the adjoint system, whose solution ψ satisfies

$$(I - \partial_x F)^\top \psi = \nabla_x \ell(x^\star), \qquad \nabla f(\theta) = \partial_\theta F^\top \psi,$$

has one right-hand side per readout. Forward mode is preferable when the same equilibrium and mask parameterization serve several scalar readouts, as in the node-level study, since one solve of

Equation (D.2) then supplies the gradients of all of them. Adjoint mode is cheaper when there is one readout per equilibrium, as in the graph-classification study. The experiments computed the endpoint gradients in forward mode. Recomputed in adjoint mode on all 480 graph-classification targets, they agree with the forward ones to relative 9.1×10^{-10}, so no reported number depends on the route. Counted in iterations times right-hand sides and in units of one cold-start coalition solve, the two endpoints together cost 2.07 in adjoint mode and 15.5 in forward mode (medians over targets), which is the basis of the main paper's "about two coalition evaluations".

Equal Boolean-query budgets do not imply equal total model cost, and the comparisons in the main paper are stated in terms of coalition queries.

D.4 The Mask in the Two Model Families

In the graph-classification study the groups are node sets obtained by a graph-distance Voronoi partition, the baseline is zero node features, and the model is a fixed-point layer

$$H = F(H, \theta) = \sigma\big(\mathsf{S}\, H\, \Theta_{\mathrm{rec}} + \Xi_X(\theta)\big), \tag{D.3}$$

with S the graph propagation operator, Θ_{rec} the recurrent weight matrix, Ξ_X the input drive and σ the SiLU nonlinearity applied componentwise. The mask acts on the node features entering Ξ_X, on which the drive depends affinely, so $\theta \mapsto \Xi_X(\theta)$ is affine.

These models satisfy Equation (D.1). Differentiating Equation (D.3) in H gives $\partial_H F[\Delta H] = \sigma' \odot (\mathsf{S}\, \Delta H\, \Theta_{\mathrm{rec}})$, with σ' evaluated entrywise at the pre-activation. Vectorizing turns the linear part into $\Theta_{\mathrm{rec}}^{\top} \otimes \mathsf{S}$, whose spectral norm is $\|\Theta_{\mathrm{rec}}\|_2 \, \|\mathsf{S}\|_2$, and the elementwise factor into a diagonal matrix of operator norm at most $\|\sigma'\|_\infty$. Hence

$$\|\partial_H F\|_{\mathrm{op}} \le \|\sigma'\|_\infty \cdot \|\mathsf{S}\|_2 \cdot \|\Theta_{\mathrm{rec}}\|_2 \,.$$

The parameterization caps the recurrent weight's spectral norm at 0.8, the normalized propagation operator has spectral norm one, and $\|\sigma'\|_\infty < 1.1$ for the SiLU, so the bound is $1.1 \times 1.0 \times 0.8 = 0.88$. The pipeline records the three factors and their product for every trained model, each maximized over its evaluated graphs. The recorded norms match their nominal values to floating-point roundoff.

In the node-level study the groups are feature groups of a trained graph convolutional model, the baseline is either zero or the feature mean over the graph according to the configuration, and the mask again acts affinely on the feature columns. Architectures, training, radii, baselines and grouping rules are specified in Appendix E.

E Experimental Protocol

This appendix specifies the two studies, the sampled designs, the controls, the estimand and the comparison with the official OddSHAP implementation, in enough detail to reconstruct any reported number from the frozen artifact.

E.1 Graph Classification

The primary study uses MUTAG, PROTEINS, NCI1 and IMDB-BINARY from the TUDataset collection (Morris et al., 2020), the four benchmarks most often reported in graph-classification explanation work. They span molecular, protein and social graphs, so the four datasets do not share one graph statistic. Each dataset is split once per model seed by a deterministic stratified rule. Within each label the n graphs carrying it are permuted by the seed, the first $\lfloor 0.70n \rfloor$ go to training, the next $\lfloor 0.15n \rfloor$ to validation and the rest to test, with each of the first two counts at least one. The three lists are then permuted again. Five seeds give five independently trained models per dataset.

From the frozen permutation of its test split, each model contributes the first 24 graphs with at least eight nodes. PROTEINS is the only dataset where this skips graphs, two to four per model. The result is 480 model-graph evaluation units. Because different seeds may draw the same graph into their test splits, these units carry 440 distinct dataset-qualified graph identifiers, namely 94 for MUTAG, 114 for PROTEINS, 119 for NCI1 and 113 for IMDB-BINARY. The bootstrap nests each graph within

its model seed, so a graph appearing under two seeds sits in two seed-specific clusters and is not linked across them.

Each model is the fixed-point layer of Equation (D.3) with 32 hidden units, SiLU activation, a recurrent spectral-norm cap of 0.8, and damping 0.8 during training. Training runs for at most 180 epochs with patience 35, batch size 32, learning rate 0.01 and weight decay 5×10^{-4}. Attribution uses the equivalent undamped fixed-point equation, solved to tolerance 10^{-10}, and the linearized system Equation (D.2) is solved to the same tolerance. Appendix D verifies the contraction hypothesis for these models.

Both studies ran on one workstation, an AMD Ryzen 9 7950X with 64 GB of memory and an NVIDIA RTX 4090, under Windows 11. Training uses the GPU when one is present. Every stage after it, the coalition evaluations, the endpoint solves, the regressions and the bootstrap, is CPU-only, and so is the released artifact. Package versions are pinned in the artifact.

The eight groups of a graph are its graph-distance Voronoi cells. Distances are unweighted shortest-path distances after removing self-loops. The first center is a node of maximum degree, ties broken by the smaller index. Each further center maximizes the distance to the centers already chosen, with ties broken first by larger degree and then by smaller index. Every node is assigned to its nearest center, with the order in which centers were chosen breaking distance ties. The construction is purely combinatorial and involves no learned clustering. The masked-out groups take a zero node-feature baseline, and the explained scalar is the logit of the class predicted under the full input, not the logit of the true class.

E.2 The Node-Level Study

The supporting study explains node predictions on the three citation networks Cora, CiteSeer and PubMed (Sen et al., 2008) under the standard Planetoid splits (Yang et al., 2016). Each checkpoint is a tied recurrent graph convolutional model with 32 hidden units, SiLU activation, damping $\alpha = 0.8$, a recurrent spectral-norm cap of 0.7 and 16 unrolled steps during training, optimized by AdamW with learning rate 0.01 and weight decay 5×10^{-4} for at most 240 epochs with patience 50 on validation accuracy. Every coalition equilibrium is solved to tolerance 10^{-7}, and so are the two linearized systems.

Targets are chosen without reference to the labels. Test nodes are partitioned by the class predicted under the full input, each block is permuted by the configuration seed offset by 7001, and 64 targets are drawn round-robin across the blocks, so the predicted classes are represented as evenly as the test split allows. The explained scalar is again the predicted-class logit.

Nine baseline configurations, three datasets by three trained seeds, use eight balanced random feature groups and a zero baseline, ten budgets $48, 64, 80, 96, 104, 112, 120, 128, 144, 160$ and 100 paired repetitions. Balanced random grouping permutes the feature coordinates with a seeded permutation and assigns positions $i, i + 8, i + 16, \ldots$ to group i, so the group sizes differ by at most one.

Seventeen further configurations run over seven budgets $48, 64, 80, 96, 112, 128, 160$ and 40 paired repetitions. Nine complete a two-by-two grid over grouping and baseline. The baseline configurations occupy the corner with balanced random groups and a zero baseline, and for each dataset the other three pair activation-and-weight feature clusters with a mean baseline, the same clusters with a zero baseline, and balanced random groups with a mean baseline. Two move the recurrent norm cap of the Cora model, to 0.5 and to 0.9. The remaining six change the stopping tolerance of an auxiliary fixed-point solver, between 10^{-5} and 10^{-7}.

That solver benchmarks warm-start strategies and does not produce a coalition value. Every coalition equilibrium comes from the tolerance fixed at 10^{-7} above, in those six configurations as in the rest. Their frozen games agree with the baseline to 5×10^{-6}, which is single-precision noise, so what they report is how the comparison survives that noise on the same game rather than sensitivity to a protocol choice. None of the seventeen configurations is an additional dataset, and the main paper reports the nine baseline ones only.

Activation-and-weight clustering describes each feature by the concatenation of a seeded random projection of its node-value profile with its normalized learned input-weight column, then runs farthest-point initialization followed by at most 40 Lloyd updates, repairing an empty cluster de-

terministically. A mean baseline replaces an absent coordinate by its mean over the nodes of the graph.

E.3 Designs and Sampling

The graph-classification sweep runs eight budgets, $48, 64, 80, 96, 112, 128, 144$ and 160, at 50 paired repetitions each. Of these, 48, 64 and 80 are the budgets whose interior designs satisfy $m < 91$, and the confirmatory analysis is restricted to them for the reason the main paper gives.

Interior coalitions are drawn by the official PolySHAP sampler (Fumagalli et al., 2026b) at a pinned commit, without replacement and in complementary pairs, so that selecting S also selects $N \setminus S$. The sampler returns a nonnegative weight for each drawn row, and W is the diagonal square root of those weights. The analysis in this document uses no property of the sampler beyond $W \succ 0$. The empty and full coalitions are not sampled rows. They enforce the efficiency constraint exactly and are two of the B evaluations.

On a design of this kind the efficiency-constrained quadratic estimator $\mathcal{A}_2$ is second-order PolySHAP, and by the pairing theorem of Fumagalli et al. (2026b) it coincides with paired KernelSHAP wherever the sampled degree-two design has full column rank.

Pairing itself limits that rank. Center the cube at $\widetilde{y} = y - \mathbf{1}/2$, so that a complementary pair sits at $\widetilde{y}$ and $-\widetilde{y}$. A coefficient vector in $\mathrm{range}(P_2)$ is a polynomial vanishing at both endpoints, and splitting it into its even and odd parts in $\widetilde{y}$ gives subspaces of dimension 28 and 7. The sum and the difference of a pair's two rows probe the even and the odd part separately, so $p = m/2$ pairs give

$$\mathrm{rank}(A_2) \le \min(p, 28) + \min(p, 7).$$

At $B = 48$ the bound is 30, below the 35 that full column rank would require, so the equivalence theorem does not apply there. From $B = 64$ on the bound is not binding. We therefore keep the paper's own name for $\mathcal{A}_2$ in the tables below.

Where masks were not stored, the design is regenerated from

$$\text{random state} = s_{\text{design}} + 1009\, B + j_{\text{rep}}, \tag{E.1}$$

where s_{design} is the seed archived with the run and j_{rep} is the zero-based repetition index. When the node-level study has no archived seed it uses $1701 + s_{\text{configuration}}$, with $s_{\text{configuration}}$ the integer seed naming that configuration. A regeneration is accepted only when it is bitwise equal to every archived mask array available for that cell. Within one comparison all in-house methods receive the same regenerated design, the same coalition values, the same W and the same target columns.

E.4 The Four Controls

Write κ for the vector of endpoint sensitivity magnitudes of a target and w for the weights Equation (4) produces from it. Three controls replace the vector that enters that map and change nothing else, so the estimator itself is untouched. The *uniform-weight control* sets $w_i = 1/g$ directly. The *group-size control* replaces κ_i by the number of nodes or feature coordinates in group i and sends that through Equation (4). The *label-permutation control* keeps w and permutes its coordinates. The *cubic-ridge control* instead omits the supplied block on every target, running the generalized ridge on the cubic space residualized against the quadratic model alone. This is the omitted branch of Appendix C, taken by decision rather than by the retention test. It removes the selected direction together with the separate penalty attached to it, so it bounds what generic cubic regularization achieves rather than isolating basis selection by a single factor.

Let

$$\mathcal{I} = \{\pi \in \mathfrak{S}_8 :\ \pi^2 = \mathrm{id},\ \pi(i) \neq i \text{ for all } i\}$$

be the permutations used by the label-permutation control. Its elements are the perfect matchings of eight players, so $|\mathcal{I}| = 7 \cdot 5 \cdot 3 \cdot 1 = 105$, and they are exactly the elements of $\mathfrak{S}_8$ of cycle type 2^4, hence a single conjugacy class, which means

$$\sigma\, \mathcal{I}\, \sigma^{-1} = \mathcal{I} \qquad \text{for every } \sigma \in \mathfrak{S}_8. \tag{E.2}$$

For $\pi \in \mathcal{I}$ the control supplies the weights $w \circ \pi$, where $(w \circ \pi)_i = w_{\pi(i)}$. Write $\widehat{\phi}(w)$ for what the algorithm returns from a fixed design and target when it is given the weights w. On one target, one mask design and one repetition the control reports

$$\frac{1}{|\mathcal{I}|} \sum_{\pi \in \mathcal{I}} \big\| \widehat{\phi}(w \circ \pi) - \phi(v) \big\|_2^2, \tag{E.3}$$

and that value is what enters the hierarchy of Appendix E.5.

The quantity Equation (E.3) is invariant under relabeling the players. Let $\sigma \in \mathfrak{S}_8$ act at once on the masks, the polynomial columns, the Shapley map and the weights, writing $(\sigma \cdot w)_i = w_{\sigma^{-1}(i)}$. Comparing coordinates gives

$$(\sigma \cdot w) \circ \pi = \sigma \cdot \big(w \circ \sigma^{-1} \pi \sigma\big),$$

since both sides send i to $w_{\sigma^{-1}\pi(i)}$. Part (iv) of the guarantee theorem makes the estimator equivariant, so on the relabeled problem the summand at π is

$$\big\| \sigma \cdot \widehat{\phi}(w \circ \sigma^{-1} \pi \sigma) - \sigma \cdot \phi(v) \big\|_2^2 = \big\| \widehat{\phi}(w \circ \sigma^{-1} \pi \sigma) - \phi(v) \big\|_2^2,$$

since a permutation preserves the Euclidean norm. As π runs over $\mathcal{I}$ so does $\sigma^{-1} \pi \sigma$, by Equation (E.2), and the sum is unchanged. A control built from a proper subset of $\mathcal{I}$, or from any set not closed under conjugation, would lose this guarantee.

E.5 Estimand and Hierarchical Bootstrap

Let $\mathrm{SSE}_{d,b,a}$ be the squared attribution error of algorithm a on dataset d at budget b, summed over model seeds, graphs, paired repetitions and the eight coordinates, and let quad be the efficiency-constrained quadratic estimator. The balanced relative error and the gain over a reference are

$$T_a = \sum_d \sum_b \frac{\mathrm{SSE}_{d,b,a}}{\mathrm{SSE}_{d,b,\mathrm{quad}}}, \qquad G(A/R) = \frac{T_R}{T_A}, \tag{E.4}$$

the sums running over the four datasets and the three confirmatory budgets. Normalizing each cell by the quadratic error is what makes the twelve cells commensurable. Without it a dataset whose logits are larger would dominate the total. In particular $T_{\mathrm{quad}} = 12$.

A draw resamples datasets with replacement, and within each drawn dataset it resamples model seeds. Graphs and repetitions are then resampled crosswise rather than one inside the other. A drawn seed receives one multinomial over its graphs and, at each budget, one multinomial over its repetitions, and the two weight the same error tensor. The eight coordinates of a unit move together, so no attribution vector is broken apart, and budgets are fixed conditions rather than resampled ones. Within an in-house comparison every method receives the same drawn indices, which is what makes the ratio of two totals a paired quantity. The normalizers $\mathrm{SSE}_{d,b,\mathrm{quad}}$ stay at their point values, so a draw perturbs the numerators only. The reported interval is the empirical 2.5% and 97.5% quantiles of the resulting gains over $100{,}000$ draws, and the reported $P(G > 1)$ is the fraction of draws whose gain exceeds one. Both summarize one comparison at a time. No simultaneous correction across the four is applied, and the fraction is a draw frequency, not a classical p-value.

The base seed is 20260722. Resampling inside the dataset at position p uses $20260722 + 1000003\,p$, and the draw across datasets uses $20260722 + 700001$. The mixture parameter $\varepsilon = 0.05$, the unit-norm cubic coefficients, the relative retention threshold 10^{-12}, the candidate grid, the four controls, the 105 involutions, the averaging order of Equation (E.3), the bootstrap hierarchy and these seeds were fixed before any error was computed. None of them, and none of the training settings above, was tuned. No hyperparameter search was run at any stage, so no value reported here was selected against a measured error.

E.6 The Official OddSHAP Comparison

The comparison with OddSHAP (Fumagalli et al., 2026a) runs the authors' implementation in `shapiq` 1.6.0, pinned to a single commit, on the same four datasets, the same trained models, the same graphs, the same eight groups and the same three budgets. It differs from the in-house comparisons in three respects, each recorded in the frozen summary. Each method uses its native sampler,

OddSHAP drawing uniformly over interior coalition sizes and ours using the paired PolySHAP design, so the masks are not paired across methods. OddSHAP contributes 10 repetitions per model seed against our 50, and Appendix F reports the sensitivity of every headline number to using only our first ten. Our estimator additionally consumes the two endpoint gradients. Our runs computed them in forward mode, two linearized fixed-point solves with g right-hand sides each, and Appendix D gives the one-right-hand-side adjoint alternative and both measured costs.

The estimand differs accordingly, and is not Equation (E.4). Because the repetition counts are unequal, each dataset-budget cell reports the squared error summed over model seeds, graphs and coordinates, then averaged over that method's own repetitions. The twelve cells are divided by the OddSHAP value rather than the quadratic one and weighted equally. The bootstrap draws datasets, model seeds and graphs once and shares them between the two methods, resampling each method's repetitions within its own count. The resulting interval is descriptive, since this comparison is secondary evidence.

What the protocol equalizes is the number of Boolean coalition queries. The frozen verification records that the official run issues exactly B coalition evaluations at every budget, and that the ground-truth Shapley vectors the two pipelines are scored against agree exactly. Reported errors are therefore comparable at a matched coalition budget, which is the quantity the main paper claims.

F Complete Results and Secondary Analyses

This appendix gives the results behind the experimental claims of the main paper, together with the analyses kept out of it. Gains are the ratio Equation (E.4) of balanced relative errors. Every value below carries four significant digits, except the bootstrap frequencies $P(G > 1)$, which carry four decimal places.

F.1 The Primary Comparisons in Detail

Throughout this subsection the entry for the label-permutation control is the involution-averaged error Equation (E.3), the quantity that enters the hierarchy. Entries for the other methods are the errors of single estimates.

Table F.1 resolves the aggregate by dataset, and Table F.2 resolves it further into the twelve summands of Equation (E.4). Reading it by dataset shows where the methods separate. On MUTAG, PROTEINS and NCI1 no two of them differ by as much as a fifth of the quadratic error in any cell. On IMDB-BINARY the proposed estimator and the group-size control fall to between 0.006 and 0.010 of that reference, while uniform and permuted weights reach only 0.072 to 0.321. The aggregate intervals inherit their upper ends from those three cells.

Withholding one dataset at a time gives Table F.3. The four gains stay above one throughout. Removing IMDB-BINARY is the informative case, since it is the dataset that separates the methods most, and the four gains there are 1.290, 1.136, 1.174 and 1.177.

F.2 Geometry of the Selected Direction

The retention test of the main paper accepted the supplied regressor in all $72{,}000$ low-budget units, so the estimator never fell back to the omitted branch on this evidence. Table F.4 summarizes how far the selected direction sits from the symmetric one. The weights are strictly positive by construction and stay close to uniform in the typical unit, at a median entropy of 1.912 against $\log 8 = 2.079$. After residualization the sine of the principal angle between the sensitivity-informed and the uniform regressor has median 0.3535, and the corresponding Shapley corrections 0.3397, so the weights change the direction in the sampled design, not merely its scale.

F.3 The Node-Level Study in Full

Table F.5 gives the node-level study in full. The proposed estimator has the lowest pooled error in every family. Three of the rows hold one stage of the estimator fixed instead of replacing the weights, and they locate what each stage contributes. Partial regression alone stops once the coefficient of the supplied regressor is estimated, with no ridge term and no shrinkage. Its ratio of 2.030 on the baseline

Dataset	Control	Gain	95% interval	$P(G > 1)$
MUTAG	label permutation	1.380	[0.9886, 1.807]	0.9557
	uniform weights	1.253	[0.9727, 1.568]	0.9121
	group size	1.249	[1.011, 1.510]	0.9920
	cubic ridge	1.270	[0.9766, 1.598]	0.9227
PROTEINS	label permutation	1.197	[1.002, 1.465]	0.9796
	uniform weights	1.046	[0.8467, 1.238]	0.6653
	group size	1.196	[1.016, 1.359]	0.9999
	cubic ridge	1.158	[0.9721, 1.513]	0.9402
NCI1	label permutation	1.261	[0.9960, 1.397]	0.9564
	uniform weights	1.070	[0.9728, 1.170]	0.8251
	group size	1.059	[1.005, 1.091]	0.9913
	cubic ridge	1.078	[0.9784, 1.182]	0.8549
IMDB-BINARY	label permutation	28.99	[6.703, 56.61]	1.0000
	uniform weights	18.07	[4.664, 34.31]	1.0000
	group size	1.234	[1.062, 1.576]	0.9982
	cubic ridge	20.29	[5.610, 38.18]	1.0000

Table F.1: Gains within each dataset, aggregated over its three confirmatory budgets, with marginal percentile intervals from 100,000 hierarchical draws over model seeds, graphs and repetitions. All sixteen point gains exceed one. Eight of the sixteen intervals have a lower endpoint below one. These are descriptive, and the predeclared comparison is the aggregate of the main paper. $P(G > 1)$ is the fraction of draws with gain above one and is not a classical p-value. The cubic-ridge control drops the supplied block together with the penalty attached to it, so it is not a single-factor isolation of basis selection.

Dataset	B	Proposed	Label perm.	Uniform	Group size	Cubic ridge	Quadratic SSE
MUTAG	48	0.3748	0.5714	0.5023	0.4956	0.5172	0.004897
	64	0.3058	0.4004	0.3695	0.3726	0.3694	0.002849
	80	0.2261	0.2791	0.2646	0.2646	0.2644	0.001914
PROTEINS	48	0.2510	0.3241	0.2712	0.3162	0.2763	0.3472
	64	0.2169	0.2508	0.2220	0.2560	0.2560	0.2093
	80	0.1835	0.2046	0.1884	0.2069	0.2219	0.1373
NCI1	48	0.2813	0.3867	0.3085	0.3021	0.3157	1.209
	64	0.2405	0.2963	0.2571	0.2545	0.2576	0.7551
	80	0.1963	0.2223	0.2026	0.2041	0.2010	0.4851
IMDB-BINARY	48	0.008380	0.3211	0.1954	0.01010	0.2053	17.49
	64	0.006837	0.1832	0.1172	0.008586	0.1335	10.83
	80	0.006043	0.1121	0.07152	0.007559	0.09257	6.710
T_a		2.298	3.552	2.970	2.699	3.111	12

Table F.2: The twelve summands of Equation (E.4), each the squared attribution error of one method divided by that of the efficiency-constrained quadratic estimator in the same cell. The divisor itself is in the last column. Summing the unrounded entries of a column gives the balanced relative error, and the ratio of two totals is the gain reported in the main paper.

Omitted	Label perm.	Uniform	Group size	Cubic ridge
MUTAG	1.655	1.319	1.126	1.409
PROTEINS	1.684	1.390	1.166	1.432
NCI1	1.676	1.394	1.227	1.479
IMDB-BINARY	1.290	1.136	1.174	1.177
All four	1.546	1.293	1.175	1.354

Table F.3: Point gains recomputed with one dataset withheld. The last row repeats the four-dataset aggregate for reference. The smallest entry is 1.126.

Quantity	Min	25%	Median	75%	Max
Residual energy	0.02222	0.02974	0.03135	0.03284	0.04579
Weight entropy	0.7809	1.801	1.912	1.973	2.074
Distance to unif.	0.03732	0.1643	0.2058	0.2962	0.7554
Residual sine	0.05507	0.2885	0.3535	0.4174	0.7286
Correction sine	0.01419	0.2389	0.3397	0.4600	1.000

Table F.4: Diagnostics over all 72,000 low-budget units, with the first and third quartiles in the middle columns. Relative residual energy is δ divided by the sampled energy of q, and distance to uniform is $\|w - \mathbf{1}/g\|_2$. The last two rows give the sine of the principal angle between the sensitivity-informed and the uniform direction, after quadratic residualization and in the induced Shapley correction respectively. Entropy is in nats, with $\log 8 = 2.079$ at the uniform weights.

	Baseline (9)	Grouping (9)	Replicate (6)	Norm cap (2)
Quadratic estimator	26.95	2.269	32.25	7.362
Cubic ridge	3.004	1.098	4.107	1.705
Uniform, fixed γ	2.842	1.099	3.875	1.495
Uniform weights	2.813	1.094	3.847	1.513
Partial regression alone	2.030	1.719	1.759	1.326
Sensitivity, fixed γ	1.035	1.009	1.032	1.011

Table F.5: Pooled squared error divided by that of the proposed estimator, by configuration family, with the number of configurations in each. The nine baseline configurations are the main paper's node-level evidence. *Grouping* fills the rest of the two-by-two grid over grouping and baseline, *Norm cap* moves the recurrent cap, and *Replicate* rebuilds three baseline games under an auxiliary solver setting that does not enter them, so it repeats the same comparison at the single-precision noise level. Pooling is over budgets, repetitions, targets and coordinates within a family. *Partial regression alone* estimates the coefficient of the supplied regressor and stops. *Fixed* γ keeps the supplied-regressor shrinkage at one instead of selecting it.

family is what the generalized-ridge stage adds there. Fixing the supplied-regressor shrinkage at $\gamma = 1$ instead of selecting it costs about 0.03 either way, 1.035 against 1 with the sensitivity-informed direction and 2.842 against 2.813 with uniform weights. What separates those two pairs is the direction, not the shrinkage. The three references keep their order across all four families, and only the partial-regression ablation moves, from below the uniform controls on the baseline family to above them under a changed grouping.

F.4 Noisy Sampled Responses

Table F.6 perturbs the sampled interior responses of the nine baseline configurations at three of their ten budgets, by additive Gaussian, Student t_3 and coordinate-weighted Gaussian noise at four levels, with 50 repetitions on 32 targets in each configuration. The proposed estimator has the lower pooled error in 38 of the 39 cells. The median ratio is 1.018 and the largest is 2.333. The single exception is 0.9998, at the largest budget and the largest level of the weighted perturbation. Down each perturbation block the ratio contracts toward one, so the advantage attenuates as the perturbation grows. The endpoint sensitivities are held clean throughout, so nothing here speaks to the effect of noisy derivatives.

F.5 The Empty-Endpoint Control

The weight rule reads both endpoint gradients. A control that builds the same weights from the empty-mask gradient alone, through $\kappa_i = |K_{0i}|$ and otherwise the identical pipeline, has 1.006 times the balanced relative error of the two-endpoint rule, with interval $[0.9906,\ 1.034]$ and 0.7641 of the draws above one. The same ratio stays between 1.004 and 1.010 with any one dataset withheld. Performance is therefore insensitive to whether the weights are built from the empty gradient alone or from both. We report the two-endpoint rule because it was the predeclared choice, because it

Perturbation	Level	$B = 96$	$B = 128$	$B = 160$
none	0	2.333	1.779	1.955
Gaussian	10^{-4}	2.280	1.703	1.626
	10^{-3}	1.279	1.094	1.013
	10^{-2}	1.014	1.007	1.000
	0.05	1.002	1.001	1.000
Student t_3	10^{-4}	2.277	1.691	1.644
	10^{-3}	1.270	1.083	1.018
	10^{-2}	1.010	1.005	1.003
	0.05	1.003	1.001	1.002
Weighted Gauss.	10^{-4}	2.301	1.732	1.786
	10^{-3}	1.401	1.164	1.078
	10^{-2}	1.015	1.011	1.005
	0.05	1.001	1.001	0.9998

Table F.6: Pooled error of the uniform-weight control divided by that of the proposed estimator, on the nine baseline node-level configurations with the sampled interior responses perturbed. The endpoint sensitivities are not perturbed, so this measures robustness of the estimation stage and not of the selection stage.

Dataset	B	OddSHAP	Proposed	Ratio
MUTAG	48	5.534	3.671×10^{-5}	1.508×10^{5}
	64	7.075×10^{-1}	1.743×10^{-5}	4.060×10^{4}
	80	4.229×10^{-5}	8.657×10^{-6}	4.885
PROTEINS	48	1.658	1.742×10^{-3}	951.8
	64	2.331×10^{-1}	9.079×10^{-4}	256.7
	80	3.068×10^{-3}	5.040×10^{-4}	6.087
NCI1	48	1.776×10^{1}	6.801×10^{-3}	2612
	64	2.363	3.632×10^{-3}	650.6
	80	1.098×10^{-2}	1.905×10^{-3}	5.766
IMDB-BINARY	48	4.842×10^{1}	2.932×10^{-3}	1.652×10^{4}
	64	1.137×10^{1}	1.480×10^{-3}	7681
	80	1.530×10^{-1}	8.110×10^{-4}	188.6

Table F.7: Squared attribution error at matched Boolean coalition budgets, summed over model seeds, graphs and coordinates and averaged over each method's own repetitions, with each method using its native sampler. Ratios range from 4.885 to 1.508×10^{5} with median 801.2.

treats the two endpoints symmetrically, and because a magnitude built from both cannot cancel when a group responds strongly but with opposite signs at the two ends.

F.6 The Official OddSHAP Comparison in Full

Table F.7 gives the twelve cells. The proposed estimator has the lower error in all of them, and in all 60 constituent model-seed-budget cells, where the smallest advantage is a factor of 2.262 in repetition-averaged squared error. Normalizing each cell by the OddSHAP error and weighting the twelve equally gives 21.63, with a descriptive hierarchical interval of $[13.77,\ 81.87]$ from $100{,}000$ draws. Repeating the aggregation with only the first ten of our fifty repetitions, so that both methods contribute ten, gives 21.41 and leaves all 12 and all 60 comparisons unchanged, so the unequal repetition counts do not carry the result. The ratio drops by between one and four orders of magnitude from budget 64 to budget 80 in every dataset. The reference error falls very steeply across that step while ours falls smoothly, so the comparison is at its least favorable to us at the top of the reported range.

Name in this paper	Frozen identifier
the proposed estimator (node-level study)	`EndpointSensitivityJARF` `GuidedJARF`
label-permutation control	`MisalignedAverage`
uniform-weight control	`UniformJARF`
group-size control	`GroupSizeJARF`
cubic ridge without a supplied regressor	`DARF`
efficiency-constrained quadratic estimator	`Direct2`
empty-endpoint control	`OneEndpointJARF`
partial regression alone	`GuidedAHD`
sensitivity basis, γ fixed	`GuidedHARF`
uniform basis, γ fixed	`UniformHARF`

Table G.1: Paper terminology and the identifiers carried by the frozen result files. The two studies name the proposed estimator differently. The last three rows appear only in Table F.5 and the empty-endpoint row only in Appendix F.5.

F.7 Numerical Checks

Efficiency is an identity of the estimator and holds to rounding. Over the 20 graph-classification runs, one per dataset and model seed, the largest observed $\left|\mathbf{1}^\top \widehat{\phi} - d_v\right|$ is 1.954×10^{-14}, at the level double precision allows.

The admissibility hypothesis of the guarantee theorem needs no measurement here. The candidate grid contains $(\gamma, \lambda) = (0, \infty)$, whose smoother is H_2 and whose degrees of freedom are therefore $\operatorname{rank}(A_2)$. Since $A_2 = WX_2P_2$ and P_2 projects onto $\mathbf{1}^\perp$ in a space of 36 nonconstant monomials of degree at most two, that rank is at most 35, while every design at the confirmatory budgets has $m \geq 46$ interior rows. Hence $m - \operatorname{rank}(A_2) \geq 11 > 0$ in every unit of every study reported here, so an admissible candidate always exists.

G Reproducibility and Implementation Mapping

The evidence behind this paper is available at

`https://anonymous.4open.science/r/endpoint-sensitivities-shapley-artifact`

This appendix connects the two, showing what the frozen files call the paper's names, where each result is stored, and how to rebuild the tables.

G.1 Implementation Labels

The frozen arrays predate the paper's terminology, so the identifiers in them differ from the names used above. Table G.1 is the whole translation.

The identifiers are compounds, and reading them as such decodes the ones this paper never reports. A prefix names the weight source, with `EndpointSensitivity` and its node-level synonym `Guided` for the endpoint sensitivities, `Uniform` for $w_i = 1/g$, `GroupSize` for group cardinality, and `OneEndpoint` for the empty-mask gradient alone. A suffix names how far the estimator runs. `JARF` runs the complete method, `AHD` stops after the partial regression, and `HARF` holds the supplied-regressor shrinkage at $\gamma = 1$ instead of selecting it. Identifiers outside Table G.1 are historical variants and name no method this paper reports.

G.2 The Frozen Evidence

One configuration governs the primary study, and its digest, which identifies the experimental state that produced every graph-classification number here, is

Reported in	File
the paper's four gains	`hierarchical_intervals.csv`
Table F.1	`dataset_intervals.csv`
Table F.2	`full_budget_descriptive.csv`
two control columns	`guided_external/*/estimates.pt`
Table F.3	`leave_one_dataset_out.csv`
Table F.4	`direction_diagnostics.json`
Appendix F.5	`summary.json`
Table F.5	`planetoid/aggregate/clean_metrics.csv`
Table F.6	`planetoid/aggregate/noise_metrics.csv`
Appendix F.7, graph classification	`regression_summary.json`
Table F.7	`dataset_budget_metrics.csv`

Table G.2: Where the results of Appendix F are stored. The first block sits in the primary study's `aggregate/` subdirectory, the second at its root, and the last in the secondary comparison. `full_budget_descriptive.csv` carries four of the six columns of Table F.2. The label-permutation and group-size columns come from the `misaligned_mean_squared_error` and `group_size` tensors of the estimate files, one directory per dataset and model seed. Appendix F.5 takes its ratio and interval from `summary.json` and its draw fraction and leave-one-out range from the first and fourth files.

```
fa3d45996bfcf99761fe3d59f9e4ffcf
e1b47c959a00724630856279dec11f6f
```

The digest is recorded in `configuration_lock.json` and repeated in the worker summaries and the aggregate, so a file that disagrees with it belongs to a different state and is not merged into these tables. Integrity of the package as a whole rests on `manifest.json`, which carries the SHA-256 digest and byte size of all 884 packaged files.

Everything reported here sits in two subtrees of `frozen_results/`, the primary study in the first and the secondary comparison in the second. They are

```
endpoint_sensitivity_guided/
official_oddshap/comparison/
```

Table G.2 locates each result within them.

The stored arrays admit an internal consistency check, because the map from endpoint data to a Hermite cube is invertible. Inverting the construction of Appendix B recovers $D(f)$, K_0 and K_1 from the cube, rebuilds the cube and its closed-form Shapley vector in double precision, and compares them with what is stored. The audit covers the 480 graph-classification payloads, stored in double precision, and the 26 node-level arrays, stored in single and carrying 1,664 target columns. The largest deviation is 1.431×10^{-5} in a cube entry and 3.974×10^{-6} in a Shapley coordinate, and every payload passes the tolerance predeclared for its dtype. The audit certifies the stored evidence, not the estimator.

G.3 Reproduction

The artifact has one entry point, `reproduce.py`. Three commands cover what a reader is likely to want, and all three are pure Python and platform independent. They are

```
python -B reproduce.py check --files-only
python -B reproduce.py tables
python -B reproduce.py oddshap
```

The first verifies every packaged digest against `manifest.json`, confirms the configuration lock and the pinned PolySHAP sources, checks the input grid, and scans all packaged bytes, compressed members included, for paths that would break anonymity. Dropping `--files-only` adds the installed package versions, a PolySHAP import, and the full OddSHAP array checks. The second reconstructs the headline and supporting summary tables from the frozen evidence and verifies them against the locked values. The third reaggregates the OddSHAP comparison from frozen, source-pinned estimates rather than rerunning the third-party estimator. All three write only under `outputs/`, and the first two finish in under a minute.

Beyond these, `smoke` runs a small estimator replay, `preflight` validates the full replay contract without executing it, and `full` runs or resumes the complete regression, aggregation and bootstrap. These reach the estimators through a checkpointed launcher and so require Windows PowerShell. Even the full replay starts from the immutable coalition games and endpoint data, so it performs no training, no equilibrium solve and no new coalition evaluation. It is CPU-parallel and takes hours.

`requirements-main.txt` pins NumPy 2.2.6, SciPy 1.15.2, scikit-learn 1.6.1 and a CPU build of PyTorch 2.6.0, under the Python 3.11 that `reproduce.py` requires. No stage of the artifact uses a GPU. Sampling uses the official PolySHAP implementation, packaged under `third_party/` with its license. The OddSHAP estimates come from an isolated Python 3.13 environment with `shapiq` 1.6.0, recorded in `requirements-oddshap.txt`. Both are pinned to commits recorded in `config/paper.json`, namely `690a0eef` for PolySHAP and `a063be30` for `shapiq`. The second environment regenerates those estimates and is needed for nothing else.

The package carries no submission author names and no affiliation identifiers, and `check` fails if any declared private filesystem path survives anywhere in it.