AUMANN-SHAP: BRIDGING SHAPLEY INTERACTION ATTRIBUTION AND INTEGRATED GRADIENTS*

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature attribution can describe a baseline-to-target change through endpoint coalitions or through a continuous path. For interacting features, these views need not divide responsibility in the same way, as endpoint Shapley splits a Harsanyi interaction pot symmetrically, whereas path attribution can reveal asymmetric roles. We introduce `Aumann-SHAP`, a discrete-to-continuous framework that preserves each endpoint interaction pot while refining how it is allocated. For every interaction set, we isolate a pure interaction surface on the local hypercube, discretize feature changes into *micro-moves*, and apply Shapley attribution to the induced micro-player game. At resolution $m = 1$, Aumann-SHAP recovers equal-split Shapley exactly; under smooth refinement, it converges to the Aumann–Shapley / straight-line Integrated Gradients allocation of the isolated interaction surface. An exact grid-state formulation requires $\mathcal{O}(k(m+1)^k)$ operations for a k-way interaction. Experiments show that transition geometry can change signs, rankings, and early edit priorities, while often leaving endpoint priorities unchanged. Thus endpoint interaction pots quantify *how much* interaction is present; Aumann-SHAP additionally captures *how responsibility develops between the endpoints*.

1 INTRODUCTION

Feature attribution can describe a baseline-to-target change from two complementary viewpoints. Cooperative-game methods such as Shapley use discrete coalition evaluations (Shapley, 1953; Lundberg & Lee, 2017), whereas path methods such as Integrated Gradients (IG) accumulate marginal changes along a continuous transition (Sundararajan et al., 2017). The former naturally exposes endpoint interactions through Harsanyi dividends (Harsanyi, 1963); the latter is closely related to the Aumann–Shapley value (Aumann & Shapley, 1974). With interacting features, however, they need not divide responsibility in the same way.

Consider the pure interaction $r(t_1, t_2) = t_1 t_2^2$ on $[0, 1]^2$. Its endpoint interaction value is one. The Boolean game therefore gives the symmetric Shapley split $(1/2, 1/2)$, while diagonal Aumann–Shapley/IG gives $(1/3, 2/3)$. The total interaction is identical; only the information used to divide it differs. This motivates our question: *can endpoint Shapley interaction attribution be refined to incorporate transition geometry while preserving the interaction quantity being allocated?*

The issue is fundamentally one of information loss. A Boolean cooperative game observes only the $2^{|u|}$ endpoint configurations of an interaction set u. Consequently, distinct predictors can induce exactly the same endpoint interaction pot ϕ_u while behaving very differently throughout the transition from $\mathbf{x}^0$ to $\mathbf{x}^1$. This distinction matters especially for counterfactual explanations: once a target input has been selected, the practical question is often not only which features differ at the endpoint, but which feature changes create, suppress, or harvest interaction effects as the model moves toward that target. Endpoint Shapley cannot distinguish such cases because the interior of the transition is never observed. We answer yes with `Aumann-SHAP`. Given endpoints $\mathbf{x}^0, \mathbf{x}^1$, we first decompose the prediction change into Harsanyi interaction pots. For each interaction set u, we isolate a pure interaction surface r_u on the local hypercube, discretize each feature movement into elementary *micro-moves*, apply Shapley to the resulting micro-player game, and aggregate micro-player payoffs back to features. The same construction interpolates between two familiar regimes: at resolution

*Code: `https://github.com/iclr-anon-2027/Aumann-SHAP`

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.

$m = 1$, it recovers equal-split Shapley; as $m \to \infty$, it converges to Aumann–Shapley/straight-line IG on each isolated interaction surface.

Figure 1 illustrates the geometric object that is missing from the endpoint game. The full cube contains the Boolean coalition evaluations used to define the Harsanyi pot, whereas an interior point $t = (t_i, t_j, t_k)$ determines an origin-anchored sub-cube whose corners encode the partially completed transition. Applying inclusion–exclusion to these interior corners yields the pure interaction surface $r_u(t)$. Thus $\phi_u = r_u(\mathbf{1})$ records the interaction at the final corner, while $r_u(t)$ retains how that same interaction develops throughout the cube. This is finer than applying IG directly to the predictor: IG returns a feature-level decomposition of $g(\mathbf{x}^1) - g(\mathbf{x}^0)$, whereas Aumann-SHAP retains the interaction index u and asks how each fixed pot ϕ_u is divided among its participants. The finite-grid construction also remains defined for non-smooth predictors. Although the induced game has km micro-players for a k-way interaction at uniform resolution m, grouping coalitions by grid state yields exact computation in $\mathcal{O}(k(m+1)^k)$, polynomial in m for fixed k.

Empirically, the refinement is selective rather than indiscriminate. On an analytically characterized nonlinear interaction it converges to the predicted path allocation; on UCI Heart Disease it changes the sign of one within-pot contribution; on German Credit it changes feature ordering in $17/138 = 12.3\%$ of successful transitions; and on a representative MNIST transition it improves the earliest counterfactual edits.

Contributions. We make three contributions: (i) a local-cube construction that preserves endpoint Harsanyi pots while exposing their interior geometry; (ii) a discrete-to-continuous theorem showing equal-split Shapley at $m = 1$ and Aumann–Shapley/IG in the refinement limit; and (iii) an exact grid-state formula together with synthetic, tabular, and image experiments showing when geometry changes attribution. Throughout, the Harsanyi pots themselves remain fixed: Aumann-SHAP changes only how each pot is allocated.

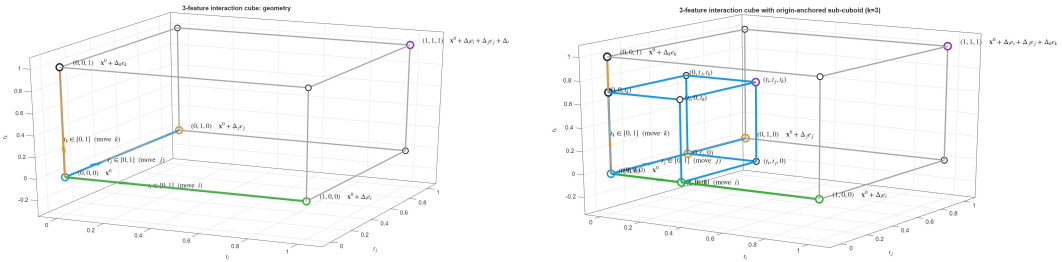
2 Related Work

Shapley-based attribution decomposes model outputs through cooperative games (Shapley, 1953; Lundberg & Lee, 2017); interaction extensions such as Shapley–Taylor and Shapley interaction indices quantify higher-order subsets (Sundararajan et al., 2020; Fumagalli et al., 2024). These methods ask how much interaction is associated with a subset. We instead condition on a fixed Harsanyi pot ϕ_u and ask how that same pot should be divided among its members once interior transition geometry is retained.

IG integrates marginal effects along a baseline-to-target path (Sundararajan et al., 2017) and is closely related to Aumann–Shapley (Aumann & Shapley, 1974); subsequent work modifies the path or its geometry (Zaher et al., 2024; Cai et al., 2025). Our continuous limit applies the path principle separately to each isolated interaction surface r_u, retaining the interaction index before aggregation. Finite-grid Aumann-SHAP also remains defined for non-differentiable models.

Integrated Hessians extends Integrated Gradients from feature effects to pairwise interactions by integrating mixed second derivatives along the baseline-to-input path (Janizek et al., 2021). It therefore shares our motivation that interaction attribution should depend on interior path geometry rather than endpoint evaluations alone. The objects are different, however: Integrated Hessians directly defines pairwise path-interaction scores for differentiable predictors, whereas Aumann-SHAP begins from an arbitrary-order endpoint Harsanyi pot ϕ_u, keeps that quantity fixed, and asks how it should be allocated among the members of u. Its finite micro-game construction additionally remains defined without derivatives and exposes equal-split Shapley and Aumann–Shapley as the two limits of the same resolution-indexed allocation.

Integrated Directional Gradients (IDG) extends path attribution from individual features to feature groups by integrating directional derivatives associated with a selected subset (Sikdar et al., 2021). More recently, H-Sets combines Hessian-based interaction discovery with a set-level IDG construction and Harsanyi dividends to attribute higher-order interaction sets in image classifiers (Mehrotra et al., 2026). These approaches are close in spirit because they also combine interaction structure with baseline-to-input geometry. Their attribution target is nevertheless different from ours: they assign an importance score to an interaction set, whereas Aumann-SHAP starts from an endpoint Harsanyi



(a) Cube corners. (b) Sub-cube to (t_i, t_j, t_k).

Figure 1: **Local slider cube for $u = \{i, j, k\}$ (a) and the sub-cube from $(0, 0, 0)$ to (t_i, t_j, t_k) (b), illustrating corner evaluations and the inclusion–exclusion construction.**

pot ϕ_u that is already fixed and asks how that same quantity should be divided among the individual members of u. The resulting resolution-indexed family additionally connects equal-split Shapley at $m = 1$ to Aumann–Shapley in the refinement limit.

Counterfactual and recourse methods primarily optimize the target input under proximity, feasibility, diversity, or causal constraints (Wachter et al., 2018; Ustun et al., 2019; Mothilal et al., 2020; Laugel et al., 2018; Karimi et al., 2021). Closest to our use of finite-change games, Lu & Borgonovo (2026) combine counterfactual analysis with Shapley effects to explain actionable changes while restricting evaluations to feasible points. Aumann-SHAP is complementary: it takes two endpoints as given and explains the transition between them. The construction therefore applies to counterfactuals but is not tied to a particular generator.

3 From Endpoint Interactions to Transition Geometry

3.1 Harsanyi Dividends and Counterfactuals

Let $N = \{1, \ldots, d\}$ index features, let $g : \mathcal{X} \to \mathbb{R}$, and fix a baseline $\mathbf{x}^0$ and target $\mathbf{x}^1$. For $S \subseteq N$, define the hybrid input

$$x_i^S = \begin{cases} x_i^1, & i \in S, \\ x_i^0, & i \notin S, \end{cases} \tag{1}$$

and the finite-change game $V(S) := g(\mathbf{x}^S) - g(\mathbf{x}^0)$, following the finite-change perspective used in counterfactual sensitivity analysis (Lu & Borgonovo, 2026), so $V(\varnothing) = 0$ and $\Delta y := V(N) = g(\mathbf{x}^1) - g(\mathbf{x}^0)$. For each nonempty $u \subseteq N$, its Harsanyi dividend is

$$\phi_u := \sum_{T \subseteq u} (-1)^{|u|-|T|} V(T), \tag{2}$$

with Möbius reconstruction $V(S) = \sum_{\varnothing \neq u \subseteq S} \phi_u$ and hence $\Delta y = \sum_{\varnothing \neq u \subseteq N} \phi_u$. Singleton pots are main effects. For $|u| \geq 2$, endpoint Shapley divides the pure pot equally, $\phi_u/|u|$, but the endpoint game alone contains no information about how that pot develops between the corners.

3.2 Local Interaction Surfaces

Fix $u \subseteq N$, $|u| = k \geq 2$, set $\delta_i = x_i^1 - x_i^0$, and let e_i be the canonical basis vector. For $t = (t_i)_{i \in u} \in [0, 1]^k$, define $\mathbf{x}_u(t) := \mathbf{x}^0 + \sum_{i \in u} t_i \delta_i e_i$, and $g_u(t) := g(\mathbf{x}_u(t))$. For $T \subseteq u$, let $t^{(T)}$ retain t_i for $i \in T$ and set the remaining coordinates to zero. The Boolean corners satisfy $\mathbf{x}_u(\mathbf{1}_T) = \mathbf{x}^T$.

Definition 1. The pure u-way interaction surface is

$$r_u(t) := \sum_{T \subseteq u} (-1)^{k-|T|} g_u(t^{(T)}), \qquad t \in [0, 1]^k. \tag{3}$$

Proposition 1. *If $t_i = 0$ for some $i \in u$, then $r_u(t) = 0$; at the far corner, $r_u(\mathbf{1}) = \phi_u$.*

All proofs are relegated to Appendix B.

Thus r_u preserves the same endpoint pot while retaining interior information. For example, $r^{(a)}(t_1, t_2) = t_1 t_2$ and $r^{(b)}(t_1, t_2) = t_1 t_2^2$ induce the same Boolean pure-interaction game and hence the same endpoint split $(1/2, 1/2)$, yet their diagonal path allocations are respectively $(1/2, 1/2)$ and $(1/3, 2/3)$. Endpoint values therefore identify how much interaction exists, not how its responsibility accumulates along a transition.

4 AUMANN-SHAP

The interaction surface r_u contains information that is absent from the Boolean endpoint game, but it is still a continuous object. We now construct a finite cooperative game that progressively exposes this interior geometry. The central idea is simple: instead of representing each feature change as one binary move from baseline to target, we divide it into smaller moves and treat those moves as players.

4.1 GRID REFINEMENT AND MICRO-PLAYERS

Fix an interaction set $u \subseteq N$ with $|u| = k \geq 2$. For every $i \in u$, choose an integer resolution $m_i \geq 1$ and write $\mathbf{m} = (m_i)_{i \in u}$. The corresponding grid of states is $\mathcal{G}_{\mathbf{m}}(u) := \prod_{i \in u} \{0, 1, \ldots, m_i\}$. A state $p = (p_i)_{i \in u} \in \mathcal{G}_{\mathbf{m}}(u)$ corresponds to the slider position $t(p) := \left(\frac{p_i}{m_i}\right)_{i \in u} \in [0, 1]^k$.

We sample the pure interaction surface on this grid: $r_p := r_u(t(p))$. By proposition 1, $r_{\mathbf{0}} = 0$ and $r_{\mathbf{m}} = \phi_u$. For a state p satisfying $p_i < m_i$, let e_i denote the unit vector in the i-th grid direction and define $\Delta_p(i) := r_{p+e_i} - r_p$. The quantity $\Delta_p(i)$ measures how much the pure interaction changes when feature i advances by one grid step, conditional on the current progress of all other participating features. We now turn these elementary movements into players.

Definition 2. For interaction set u and resolution profile $\mathbf{m}$, define the micro-player set $N_u' := \bigcup_{i \in u} (\{i\} \times \{1, \ldots, m_i\})$. Hence $|N_u'| = \sum_{i \in u} m_i$. For a micro-coalition $A \subseteq N_u'$, define its grid state $p(A) \in \mathcal{G}_{\mathbf{m}}(u)$ coordinatewise by $p_i(A) := \left|A \cap \left(\{i\} \times \{1, \ldots, m_i\}\right)\right|$. The induced cooperative game $v_u : 2^{N_u'} \to \mathbb{R}$ is $v_u(A) := r_{p(A)}$.

The micro-game preserves the interaction pot: $v_u(\varnothing) = 0$, and $v_u(N_u') = \phi_u$. More importantly, its marginal contributions have a direct geometric interpretation. For a micro-player $a = (i, s) \notin A$,

$$\begin{aligned} v_u(A \cup \{a\}) - v_u(A) &= r_{p(A)+e_i} - r_{p(A)} \\ &= \Delta_{p(A)}(i). \end{aligned} \tag{4}$$

Thus every cooperative-game marginal is exactly one local increment of the interaction surface. This identity is the key link between geometry and cooperative-game attribution. The contribution of a feature step depends on the state in which it arrives: the same feature may create little interaction when other features remain near baseline, but a large interaction once they have already moved toward their targets.

Example 1. Let $u = \{1, 2\}$ and $m_1 = m_2 = 2$. There are four micro-players, $N_u' = \{(1, 1), (1, 2), (2, 1), (2, 2)\}$. For example, the coalition $A = \{(1, 1), (2, 1), (2, 2)\}$ has state $p(A) = (1, 2)$ and therefore $v_u(A) = r_u\left(\frac{1}{2}, 1\right)$. Adding the remaining micro-player of feature 1 moves the state from $(1, 2)$ to $(2, 2)$, so its marginal contribution is $r_u(1, 1) - r_u\left(\frac{1}{2}, 1\right)$. The micro-game therefore evaluates not only whether a feature is present, but how interaction value accumulates as its movement is resolved more finely.

4.2 FEATURE-LEVEL INTERACTION ATTRIBUTION

We allocate the induced game v_u using the Shapley value (Shapley, 1953). Let Π be a uniformly random permutation of the micro-player set N_u'. For a micro-player $a \in N_u'$, let P_a^{Π} denote the set of micro-players appearing before a in Π. Its Shapley value can be written as

$$\varphi_a^{\mathrm{Sh}}(v_u) := \mathbb{E}_{\Pi}\left[v_u(P_a^{\Pi} \cup \{a\}) - v_u(P_a^{\Pi})\right]. \tag{5}$$

Algorithm 1 AUMANN-SHAP for a baseline–target transition

Require: predictor g, endpoints $(\mathbf{x}^0, \mathbf{x}^1)$, resolution $\mathbf{m}$
Ensure: feature attributions $\{S_i^{\mathrm{AS}}\}_{i=1}^d$
Construct $V(S) = g(\mathbf{x}^S) - g(\mathbf{x}^0)$ and Harsanyi pots $\{\phi_u\}$
for all $u \subseteq N$ with $|u| \geq 2$ **do**
    Construct the pure interaction surface r_u
    Sample r_u on the grid induced by $\mathbf{m}$
    Define the micro-game $v_u(A) = r_{p(A)}$
    Compute micro-player Shapley values
    Aggregate to $S_{i\to u}^{\mathrm{AS}}$ for each $i \in u$
end for
$S_i^{\mathrm{AS}} \leftarrow \phi_{\{i\}} + \sum_{\substack{u \ni i \\ |u|\geq 2}} S_{i\to u}^{\mathrm{AS}}$
return $\{S_i^{\mathrm{AS}}\}_{i=1}^d$

By Equation 4, this is equivalently $\varphi^{\mathrm{Sh}}_{(i,s)}(v_u) = \mathbb{E}_\Pi\left[\Delta_{p(P^\Pi_{(i,s)})}(i)\right]$. A micro-player therefore receives its average geometric interaction increment over all possible arrival contexts. The individual micro-player payoffs are finally aggregated back to the original feature.

Definition 3. For $i \in u$, the Aumann-SHAP share of feature i in interaction pot ϕ_u at resolution $\mathbf{m}$ is

$$S_{i\to u}^{\mathrm{AS}}(\mathbf{m}) := \sum_{s=1}^{m_i} \varphi^{\mathrm{Sh}}_{(i,s)}(v_u). \tag{6}$$

Because the Shapley value is efficient and $v_u(N'_u) = \phi_u$, the interaction shares satisfy $\sum_{i\in u} S_{i\to u}^{\mathrm{AS}}(\mathbf{m}) = \phi_u$. Hence grid refinement changes only *how* an interaction pot is divided; it never changes the pot itself. Combining the singleton effects and all non-singleton interaction shares gives the complete feature attribution for the transition:

$$S_i^{\mathrm{AS}}(\mathbf{m}) := \phi_{\{i\}} + \sum_{\substack{u\subseteq N: \\ i\in u,\ |u|\geq 2}} S_{i\to u}^{\mathrm{AS}}(\mathbf{m}), \tag{7}$$

where the resolution profile may be chosen separately for each interaction when needed. Summing over features recovers the full prediction change: $\sum_{i=1}^d S_i^{\mathrm{AS}} = \sum_{\varnothing\neq u\subseteq N} \phi_u = g(\mathbf{x}^1) - g(\mathbf{x}^0)$.

What changes under refinement? The endpoint game, the Harsanyi pots, and the quantity being explained are fixed throughout. Only the resolution at which the interior interaction geometry is observed changes. At the coarsest possible resolution, $m_i = 1$, no interior state is visible and the construction reduces to ordinary equal-split Shapley. As the resolution increases, increasingly fine geometric information enters the induced game. The next section makes both statements precise.

Algorithm 1 summarizes the complete construction from endpoint Harsanyi pots to feature-level Aumann-SHAP attributions.

By construction, $\sum_{i\in u} S_{i\to u}^{\mathrm{AS}} = \phi_u$ for every interaction pot and $\sum_i S_i^{\mathrm{AS}} = g(\mathbf{x}^1) - g(\mathbf{x}^0)$; resolution changes only how much interior geometry is used to divide each pot.

How to read the resolution? The resolution m is a fidelity parameter for the same interaction pot. At $m = 1$, the analyst accepts the endpoint symmetry assumption and obtains the usual equal split. Increasing m asks whether that symmetry persists once intermediate states are exposed. If the resulting shares remain close to their $m = 1$ values, refinement adds little; substantial changes indicate that the endpoint game was hiding asymmetric interaction geometry. Algorithm 1 therefore defines a diagnostic refinement rather than a competing attribution target: the pot ϕ_u is fixed, while the information used to divide it becomes progressively richer.

5 FROM SHAPLEY TO INTEGRATED GRADIENTS

Aumann-SHAP depends on the resolution at which the interaction surface is observed. We now characterize the two ends of this resolution scale. With one micro-move per feature, no interior

geometry is visible and the construction recovers the usual equal split of a Harsanyi interaction. Under increasingly fine refinement, the same construction converges to the continuous Aumann–Shapley / Integrated Gradients allocation of the pure interaction surface.

5.1 THE COARSE LIMIT: EQUAL-SPLIT SHAPLEY

Consider the coarsest possible discretization, $m_i = 1$ for every $i \in u$. Then each original feature contributes exactly one micro-player, so N'_u can be identified with the feature set u itself. Because the pure interaction surface vanishes whenever any participating coordinate remains at baseline (prop 1), every proper coalition of micro-players has value zero. Only the grand coalition reaches the far corner and receives the full interaction pot ϕ_u. Hence the induced game is exactly a unanimity game.

Proposition 2. *Let $u \subseteq N$ with $|u| = k \geq 2$ and suppose $m_i = 1$ for all $i \in u$. Then*

$$v_u(A) = \begin{cases} 0, & A \subsetneq N'_u, \\ \phi_u, & A = N'_u. \end{cases} \tag{8}$$

Consequently, $S^{\mathrm{AS}}_{i \to u}(\mathbf{1}) = \frac{\phi_u}{|u|}$ for every $i \in u$.

Thus the conventional equal split of an interaction pot is not a separate object in our framework. It is precisely the Aumann-SHAP allocation obtained when the transition is observed at the coarsest possible resolution.

For the example $r(t_1, t_2) = t_1 t_2^2$, the $m = 1$ game observes only $r(0,0) = r(1,0) = r(0,1) = 0$, $r(1,1) = 1$, and therefore returns $\left(\frac{1}{2}, \frac{1}{2}\right)$. The asymmetric geometry responsible for the $(1/3, 2/3)$ path allocation is invisible until interior states are introduced.

5.2 THE CONTINUOUS LIMIT

We now refine the grid. Intuitively, a random ordering of increasingly many micro-players interleaves the elementary moves of all participating features. At an intermediate fraction τ of that ordering, each feature has completed approximately the same fraction τ of its own move. The random micro-player path therefore concentrates around the diagonal $\tau\mathbf{1} = (\tau, \ldots, \tau) \in [0,1]^k$. Each elementary marginal then approaches a directional derivative of r_u along this diagonal. This yields the main discrete-to-continuous result.

Theorem 1. *Let $u \subseteq N$ with $|u| = k \geq 2$, and let $r_u \in C^1([0,1]^k)$. Consider a sequence of resolution profiles $\mathbf{m} = (m_i)_{i \in u}$ such that $\min_{j \in u} m_j \longrightarrow \infty$. Then, for every $i \in u$,*

$$\lim_{\min_{j \in u} m_j \to \infty} S^{\mathrm{AS}}_{i \to u}(\mathbf{m}) = \int_0^1 \frac{\partial r_u}{\partial t_i}(\tau\mathbf{1})\, d\tau. \tag{9}$$

Moreover, the limiting shares remain efficient: $\sum_{i \in u} \int_0^1 \frac{\partial r_u}{\partial t_i}(\tau\mathbf{1})\, d\tau = \phi_u$.

The right-hand side of equation 9 is exactly the Aumann–Shapley allocation of the interaction surface r_u along the diagonal path. Equivalently, it is straight-line Integrated Gradients applied *to the isolated interaction component* rather than directly to the full predictor.

The theorem therefore establishes the bridge announced in the Introduction:

$$\begin{aligned} m_i = 1 \quad &\Longrightarrow \quad S^{\mathrm{AS}}_{i \to u} = \frac{\phi_u}{|u|}, \\ \min_i m_i \to \infty \quad &\Longrightarrow \quad S^{\mathrm{AS}}_{i \to u} = \int_0^1 \partial_i r_u(\tau\mathbf{1})\, d\tau. \end{aligned} \tag{10}$$

The quantity being allocated is unchanged throughout: $\sum_{i \in u} S^{\mathrm{AS}}_{i \to u}(\mathbf{m}) = \phi_u$ at every finite resolution, and the same identity holds in the continuous limit. Grid refinement therefore changes only the information used to divide the pot.

Why this is not ordinary IG. Applying straight-line IG directly to g produces one total contribution per feature,

$$\mathrm{IG}_i(g) = \int_0^1 \frac{\partial g}{\partial x_i}\big(\mathbf{x}^0 + \tau(\mathbf{x}^1 - \mathbf{x}^0)\big)(x_i^1 - x_i^0)\, d\tau.$$

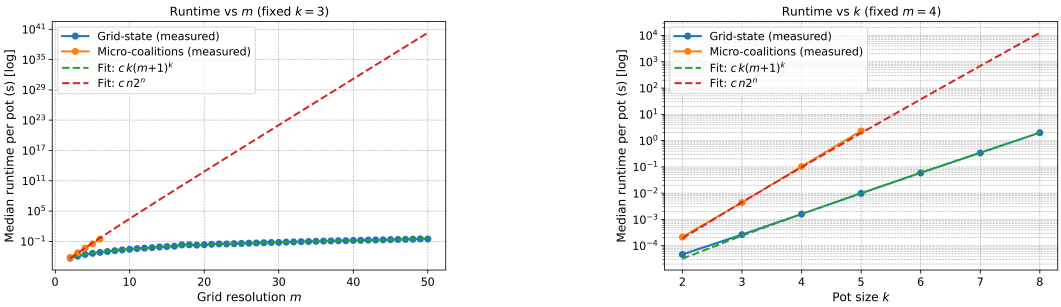


Figure 2: **Empirical scaling of exact grid-state computation.** Full micro-coalition enumeration grows exponentially in the number of micro-players, while state aggregation follows the substantially smaller grid-state complexity predicted by Prop. 3.

Aumann-SHAP instead retains the interaction index u. For every pot ϕ_u, it computes separately $S^{\mathrm{AS}}_{i\to u}$, and only afterwards aggregates these shares across interactions. Thus it answers the finer question of how a particular interaction is divided among its participants.

Non-smooth predictors. The finite-grid construction in definitions 2 and 3 does not require differentiability. It remains well defined for models such as decision trees and boosted tree ensembles because it uses only evaluations of r_u at finitely many grid states. The smoothness assumption in Thm. 1 is needed only to identify the limiting attribution with the Aumann–Shapley integral.

5.3 Exact grid-state computation

The micro-player interpretation is conceptually useful, but direct Shapley computation on v_u would be impractical. Let $n := |N'_u| = \sum_{j\in u} m_j$. Naively computing the Shapley value requires considering coalitions of n micro-players and therefore scales as $\mathcal{O}(n2^n)$.

The induced game, however, has substantial symmetry. Its value depends on a coalition A only through its grid state $p(A)$. Many distinct micro-coalitions are therefore equivalent.

For $p = (p_j)_{j\in u} \in \mathcal{G}_{\mathbf{m}}(u)$, write $|p| := \sum_{j\in u} p_j$.

Proposition 3. *For every $i \in u$,*

$$S^{\mathrm{AS}}_{i\to u}(\mathbf{m}) = \sum_{\substack{p\in\mathcal{G}_{\mathbf{m}}(u)\\ p_i<m_i}} \frac{|p|!\,(n-|p|-1)!}{n!} \times \left(\prod_{j\in u}\binom{m_j}{p_j}\right)(m_i-p_i)\Delta_p(i). \tag{11}$$

The combinatorial factors in Eq. (11) count all micro-coalitions that correspond to the same geometric state. Hence the formula performs the Shapley averaging directly over grid states rather than over individual micro-coalitions. In the uniform-resolution case $m_i = m$ for every $i \in u$, there are only $(m+1)^k$ grid states. Evaluating the exact feature shares therefore requires $\mathcal{O}\big(k(m+1)^k\big)$ operations once the residual grid has been constructed, instead of $\mathcal{O}\big(km\,2^{km}\big)$ for direct micro-coalition enumeration. Thus, for fixed interaction order k, exact computation is polynomial in the grid resolution m. The remaining exponential dependence is on interaction order, not on the number of micro-moves. This distinction is important in practice. The role of m is numerical: increasing it reveals progressively finer transition geometry. The role of k is structural: it is the order of the interaction being explained. Low-order interactions can therefore be refined substantially without creating the exponential blow-up suggested by the raw micro-player representation.

The empirical runtimes in Figure 2 mirror this distinction. As resolution m increases at fixed interaction order, direct micro-coalition enumeration rapidly becomes infeasible, whereas the grid-state computation follows the substantially milder $(m+1)^k$ state-space growth. Increasing k, by contrast, remains intrinsically costly, consistent with the exponential dependence on interaction order in Eq. (11).

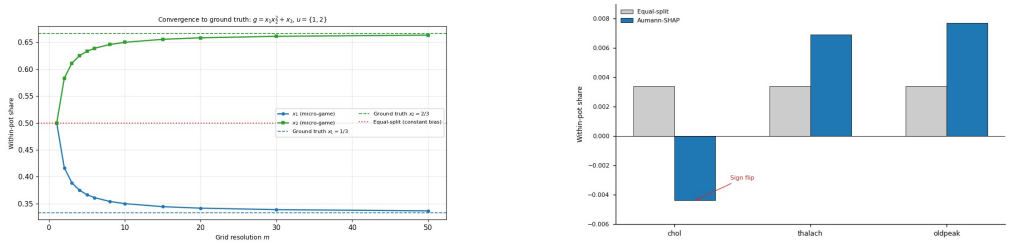


Figure 3: **Transition geometry changes interaction allocation.** Synthetic convergence to the analytical path allocation (left) and a qualitative sign change on Heart Disease (right).

6 Experiments

We test whether refinement (i) recovers an analytically known path allocation, (ii) can change a qualitative conclusion, (iii) changes feature priorities systematically, and (iv) affects sequential edit efficiency. Equal-split Shapley denotes the $m = 1$ allocation; additional implementation details are in Sec. D.

6.1 Analytical interaction

For $g(x_1, x_2, x_3) = x_1x_2^2 + x_3$ and $\mathbf{x}^0 = (0, 0, 0) \to \mathbf{x}^1 = (1, 1, 1)$, the interaction $u = \{1, 2\}$ has $r_u(t_1, t_2) = t_1t_2^2$ and $\phi_u = 1$. Equal-split Shapley gives $(1/2, 1/2)$, whereas the diagonal Aumann–Shapley allocation is $(1/3, 2/3)$. Figure 3 shows finite-resolution Aumann-SHAP converging to $(0.333, 0.667)$ by $m = 50$, while equal splitting remains fixed. This validates convergence to the analytically characterized path semantics; it does not claim that one attribution semantics is universal ground truth.

6.2 Qualitative sign change

On UCI Cleveland Heart Disease (Janosi et al., 1988) (303 observations, 13 features; XGBoost AUC 0.91), we study one transition involving $u = \{\texttt{chol}, \texttt{thalach}, \texttt{oldpeak}\}$ with $\phi_u = 0.0103$. The selected coordinates move from $(240, 171, 0.9)$ to $(258, 157, 2.6)$. Equal splitting assigns approximately $+0.0034$ to each feature, while Aumann-SHAP gives -0.0044 to `chol`, $+0.0069$ to `thalach`, and $+0.0077$ to `oldpeak`. Thus the allocation principle changes the sign of `chol`'s share inside the same fixed interaction pot.

6.3 Dataset-level priority changes

On German Credit (Hofmann, 1994) (1,000 applicants, 27 attributes), successful counterfactuals move low-scoring baselines toward a creditworthy target. For a representative Growing-Spheres transition with $\Delta V = 0.6569$ and changed features $\{X_8, X_{13}, X_{14}, X_{25}\}$, the pairwise pot $\{X_8, X_{13}\}$ is split equally as $(0.0659, 0.0659)$ by endpoint Shapley but as $(0.1143, 0.0176)$ by Aumann-SHAP at $m = 5$.

Table 1: **Representative German Credit redistribution.**

Feature	Equal split	Aumann-SHAP
X_8 (checking account)	0.0659	0.1143
X_{13} (credit history)	0.0659	0.0176

At dataset level, we aggregate 138 low-score XGBoost baselines with successful counterfactuals (mean 2.51 changed features). Aumann-SHAP and equal-split Shapley induce different strict feature-priority orderings in $17/138 = 12.3\%$ of transitions and agree in 87.7%. Geometry therefore acts as a refinement: it usually preserves endpoint priorities but changes them when within-pot asymmetry is large enough.

6.4 COUNTERFACTUAL EDIT EFFICIENCY

A ResNet-18 (He et al., 2016) trained on MNIST (LeCun et al., 2010) reaches 0.9972 test accuracy. For a representative $1 \to 7$ transition with 109 changed pixels, we rank the same editable pixels by each attribution rule and sequentially patch them from baseline to target. Aumann-SHAP uses $m = 10$ with Monte Carlo permutation sampling.

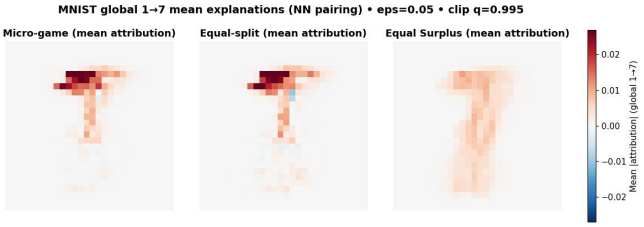


Figure 4: **Mean MNIST explanations over 200 $1 \to 7$ transitions.** Each baseline 1 is paired with its nearest target 7. Both Shapley-based methods concentrate attribution on the characteristic upper stroke, while Aumann-SHAP modifies how responsibility is distributed within the transition.

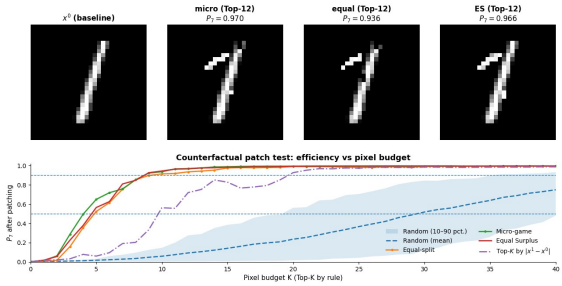


Figure 5: **MNIST counterfactual patch test.** Target-class probability after applying the top-K edits; all methods use the same editable pixels and differ only in ordering.

Table 2: **Patch-test efficiency on the representative transition.**

Method	K@0.5	K@0.9	AUC
Aumann-SHAP	**4**	**6**	**0.8928**
Equal-split	5	6	0.8916
Equal Surplus	5	7	0.8761
$\lvert x^1 - x^0 \rvert$	14	21	0.5823
Random	—	—	0.1094

Aumann-SHAP reaches $P_7 \geq 0.5$ after four edits versus five for equal-split Shapley; both reach $P_7 \geq 0.9$ after six. Their AUCs are close (0.8928 vs. 0.8916), so the observed advantage is concentrated at the earliest budgets, while magnitude ordering requires 14 edits to reach 0.5. Figure 4 provides a complementary dataset-level view over 200 nearest-neighbor $1 \to 7$ transitions: both Shapley-based methods concentrate attribution on the characteristic top bar and upper-right region, showing that Aumann-SHAP usually refines, rather than replaces, the endpoint explanation. The patch-efficiency comparison itself remains a single-transition result.

7 CONCLUSION

We introduced `Aumann-SHAP`, which preserves endpoint Harsanyi interaction pots while using transition geometry to allocate each pot among its participating features. At $m = 1$ it recovers equal-split Shapley; under smooth refinement it converges to Aumann–Shapley/straight-line IG on the isolated interaction surface; and an exact grid-state formula costs $\mathcal{O}(k(m+1)^k)$ for fixed interaction order. Synthetic, Heart Disease, German Credit, and MNIST experiments show that geometry usually refines rather than overturns endpoint attribution, but can change signs, rankings, and early edit priorities. The central distinction is therefore simple: ϕ_u answers *how much interaction is present*, while r_u contains additional information about *how responsibility for that interaction develops between the endpoints*.

AI USE STATEMENT

Generative AI tools were used to assist with research ideation, literature discovery, manuscript drafting and editing. All AI-assisted material, including theoretical claims and references were independently reviewed and verified by the authors. Generative AI was not used to generate the experimental datasets reported in this work. The authors take responsibility for the final content of the paper, including all text, mathematical claims, code, and experimental results.

ETHICS STATEMENT

Aumann-SHAP is a post-hoc attribution method for explaining the behavior of predictive models between a fixed baseline and target input. Its attributions describe the fitted model and should not be interpreted as causal effects or as recommendations for intervention. In high-stakes domains, model explanations may also inherit biases or limitations from the underlying data and predictor. Accordingly, any practical use requires domain-specific validation, appropriate feasibility and causal considerations, and independent human judgment. The Heart Disease and German Credit experiments in this paper are used only to study model-level explanation behavior; no clinical or lending decisions are made from the resulting attributions.

REPRODUCIBILITY STATEMENT

All definitions, assumptions, and theoretical results are stated explicitly, with complete proofs provided in Sec. B. Additional experimental details, including the synthetic construction, datasets, models, selected transitions, refinement resolutions, sampling procedures, and evaluation protocols, are reported in Sec. D. Code and reproducibility materials are available at `https://github.com/iclr-anon-2027/Aumann-SHAP`. The repository contains the implementation, experiment scripts, frozen non-checkpoint artifacts, SHA-256 manifests, and read-only verification tools used to reproduce the reported results.

REFERENCES

Robert J. Aumann and Lloyd S. Shapley. *Values of Non-Atomic Games*. Princeton Univ. Press, 1974.

Sylvain Béal, Eric Rémila, and Philippe Solal. Decomposition of the space of TU-games, strong transfer invariance and the Banzhaf value. *Oper. Res. Lett.*, 43(2):123–125, 2015.

Yi Cai, Thibaud Ardoin, and Gerhard Wunder. Gefa: A general feature attribution framework using proxy gradient estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6165–6192, 2025.

Célestin Chameni Nembua. Linear efficient and symmetric values for TU-games: Sharing the joint gain of cooperation. *Games Econ. Behav.*, 74(1), 2012.

Célestin Chameni Nembua and Nicolas Gabriel Andjiga. Linear, efficient and symmetric values for TU-games. *Econ. Bull.*, 3(71):1–10, 2008.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. Kernelshap-iq: Weighted least square optimization for shapley interactions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14308–14342, 2024.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *Int. Econ. Rev.*, 4(2):194–220, 1963.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pp. 770–778, 2016.

Hans Hofmann. Statlog (german credit data), 1994. UCI ML Repository.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021.

A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. UCI heart disease dataset. `https://archive.ics.uci.edu/ml/datasets/heart+disease`, 1988.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proc. FAccT*, pp. 353–362, 2021.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *IPMU*, volume 853, pp. 100–111. Springer, Cham, 2018.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. Mnist handwritten digit database, 2010. Online.

Xuefei Lu and Emanuele Borgonovo. How do i explain that decision? a new approach combining counterfactuals and sensitivity analyses. *European Journal of Operational Research*, 334(1): 200–215, 2026. doi: 10.1016/j.ejor.2026.04.035.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, volume 30, pp. 4765–4774. Curran Assoc., 2017.

Ayushi Mehrotra, Dipkamal Bhusal, Michael Clifford, and Nidhi Rastogi. H-sets: Hessian-guided discovery of set-level feature interactions in image classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pp. 607–617. ACM, 2020.

Tadeusz Radzik and Theo Driessen. On a family of values for TU-games generalizing the Shapley value. *Math. Soc. Sci.*, 65(2):105–111, 2013.

Luis M. Ruiz, Fernando Valenciano, and José M. Zarzuelo. The family of least square values for transferable utility games. *Games Econ. Behav.*, 24(1–2):109–130, 1998.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contrib. Theory Games II*, volume 28 of *Ann. Math. Stud.*, pp. 307–317. Princeton Univ. Press, 1953.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural nlp models. In *Proceedings of ACL-IJCNLP*, pp. 865–878, 2021. doi: 10.18653/v1/2021.acl-long.71.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proc. ICML*, volume 70 of *Proc. Mach. Learn. Res.*, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268. PMLR, 2020.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proc. FAT**, pp. 10–19, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018.

Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58090–58104, 2024.

A Discussion and Limitations

When geometry matters and computational scope. Aumann-SHAP refines rather than replaces endpoint attribution: in German Credit, feature ordering is unchanged in 87.7% of transitions and differs in 12.3%, with larger effects when interaction geometry is asymmetric. Exact computation costs $\mathcal{O}(k(m+1)^k)$ per k-way pot at uniform resolution m; high-order interactions therefore still require truncation, pot screening, or Monte Carlo permutation sampling (as in MNIST). The gain is that refinement is polynomial in m for fixed k, not that arbitrary high-order interactions become cheap.

Interpolation and smoothness. Finite-grid Aumann-SHAP uses only model evaluations and therefore applies to non-smooth predictors; the Aumann–Shapley/IG limit additionally assumes $r_u \in C^1$. Continuous features admit natural interpolation between endpoints. For categorical coordinates, our implementation uses an endpoint-mixture extension that preserves all Boolean corners and hence all $V(S)$ and ϕ_u, but the resulting interior geometry depends on that extension. When a domain supplies a more meaningful transition path, it should be preferred.

Causality and empirical scope. Aumann-SHAP explains a fitted model between two chosen inputs; it is not a causal effect estimator. Feasibility, immutability, and causal constraints belong to counterfactual generation (Ustun et al., 2019; Karimi et al., 2021). The Heart sign reversal is therefore model-level, not clinical advice. Results are also transition-dependent: changing either endpoint can change both pots and allocations. Finally, German Credit provides a 138-transition dataset-level analysis, whereas the current MNIST patch-efficiency result is one representative transition; broader multi-transition patch testing would strengthen the latter claim.

B Additional Theory

B.1 Proof of Proposition 1

Recall that

$$r_u(t) = \sum_{T \subseteq u} (-1)^{|u|-|T|} g_u\Big(t^{(T)}\Big).$$

Boundary vanishing. Assume that

$$t_{i_0} = 0$$

for some $i_0 \in u$.

Pair every subset $T \subseteq u$ with $i_0 \notin T$ with the subset

$$T \cup \{i_0\}.$$

Because the i_0-th coordinate of t is zero,

$$t^{(T)} = t^{(T \cup \{i_0\})}.$$

Hence the corresponding model evaluations are identical:

$$g_u\Big(t^{(T)}\Big) = g_u\Big(t^{(T \cup \{i_0\})}\Big).$$

Their inclusion–exclusion coefficients have opposite signs, since

$$(-1)^{|u|-|T \cup \{i_0\}|} = -(-1)^{|u|-|T|}.$$

The two terms therefore cancel. Pairing all subsets in this way gives

$$r_u(t) = 0.$$

Endpoint recovery. Set $t = \mathbf{1}$. For every $T \subseteq u$, masking $\mathbf{1}$ outside T gives the Boolean indicator of T, and therefore

$$\mathbf{x}_u\Big(\mathbf{1}^{(T)}\Big) = \mathbf{x}^T.$$

Consequently,

$$\begin{aligned} r_u(\mathbf{1}) &= \sum_{T\subseteq u} (-1)^{|u|-|T|} g(\mathbf{x}^T) \\ &= \sum_{T\subseteq u} (-1)^{|u|-|T|} \left[V(T) + g(\mathbf{x}^0)\right] \\ &= \sum_{T\subseteq u} (-1)^{|u|-|T|} V(T) + g(\mathbf{x}^0) \sum_{T\subseteq u} (-1)^{|u|-|T|}. \end{aligned} \tag{12}$$

The first term is precisely ϕ_u. For $|u| \geq 1$,

$$\sum_{T\subseteq u} (-1)^{|u|-|T|} = (1-1)^{|u|} = 0.$$

Thus

$$r_u(\mathbf{1}) = \phi_u.$$

□

B.2 Proof of Proposition 2

Suppose

$$m_i = 1 \qquad \text{for every } i \in u.$$

Then there is exactly one micro-player associated with each feature, and hence

$$|N'_u| = |u|.$$

Consider a proper micro-coalition

$$A \subsetneq N'_u.$$

Because A does not contain every micro-player, there exists at least one feature $i_0 \in u$ whose unique micro-player is absent. Hence

$$p_{i_0}(A) = 0.$$

By boundary vanishing,

$$v_u(A) = r_{p(A)} = 0.$$

For the grand coalition,

$$p(N'_u) = \mathbf{1},$$

and endpoint recovery gives

$$v_u(N'_u) = r_u(\mathbf{1}) = \phi_u.$$

Therefore the induced micro-game is the unanimity game

$$v_u(A) = \begin{cases} 0, & A \subsetneq N'_u, \\ \phi_u, & A = N'_u. \end{cases}$$

All $|u|$ players are symmetric in this game. By Shapley efficiency and symmetry, each receives

$$\frac{\phi_u}{|u|}.$$

Since each original feature corresponds to exactly one micro-player,

$$S^{\mathrm{AS}}_{i\to u}(\mathbf{1}) = \frac{\phi_u}{|u|} \qquad \text{for every } i \in u.$$

□

B.3 Proof of Theorem 1

Fix a feature $i \in u$. We use the random-order characterization of the Shapley value from Eq. (5).

A uniform random permutation of the micro-player set N'_u can be generated as follows. Assign independently to every micro-player $a \in N'_u$ a priority

$$U_a \sim \mathrm{Unif}[0,1],$$

and order the micro-players by increasing priority. Since the priorities are i.i.d. and continuous, ties occur with probability zero and the resulting ordering is a uniform random permutation.

Fix one micro-player

$$a = (i, s)$$

belonging to feature i, and condition on

$$U_a = q, \qquad q \in [0,1].$$

Let

$$C_j^{(a)}(q)$$

be the number of micro-players belonging to feature j that appear before a.

For $j \neq i$,

$$C_j^{(a)}(q) \sim \mathrm{Binomial}(m_j, q),$$

while for $j = i$,

$$C_i^{(a)}(q) \sim \mathrm{Binomial}(m_i - 1, q),$$

because the micro-player a itself is excluded.

Define the normalized pre-arrival state

$$Z^{(a)}(q) := \left(\frac{C_j^{(a)}(q)}{m_j} \right)_{j \in u}.$$

The marginal contribution of $a = (i, s)$ is

$$v_u(P_a^{\Pi} \cup \{a\}) - v_u(P_a^{\Pi}) = r_u\Big(Z^{(a)}(q) + \frac{1}{m_i} e_i\Big) - r_u\Big(Z^{(a)}(q)\Big).$$

Because $r_u \in C^1([0,1]^k)$, the fundamental theorem of calculus along the i-th coordinate yields

$$r_u\Big(Z^{(a)}(q) + \frac{1}{m_i} e_i\Big) - r_u\Big(Z^{(a)}(q)\Big) = \frac{1}{m_i} \int_0^1 \frac{\partial r_u}{\partial t_i} \left(Z^{(a)}(q) + \frac{\theta}{m_i} e_i \right) d\theta.$$

All m_i micro-players associated with feature i are exchangeable. Therefore, summing their Shapley values cancels the factor $1/m_i$, giving

$$S_{i \to u}^{\mathrm{AS}}(\mathbf{m}) = \int_0^1 \mathbb{E}\left[\int_0^1 \frac{\partial r_u}{\partial t_i} \left(Z^{(a)}(q) + \frac{\theta}{m_i} e_i \right) d\theta \right] dq. \tag{13}$$

We next characterize the normalized state as the grid is refined.

For $j \neq i$,

$$\mathbb{E}\left[\frac{C_j^{(a)}(q)}{m_j} \right] = q,$$

and

$$\mathrm{Var}\left(\frac{C_j^{(a)}(q)}{m_j} \right) = \frac{q(1-q)}{m_j} \leq \frac{1}{4m_j}.$$

For $j = i$,

$$\mathbb{E}\left[\frac{C_i^{(a)}(q)}{m_i}\right] = q\left(1 - \frac{1}{m_i}\right),$$

and its variance is likewise $O(m_i^{-1})$.

Hence, as

$$\min_{j \in u} m_j \longrightarrow \infty,$$

we have, for every fixed $q \in [0, 1]$,

$$Z^{(a)}(q) \xrightarrow{P} q\mathbf{1}. \tag{14}$$

Moreover,

$$\frac{\theta}{m_i} e_i \longrightarrow 0$$

uniformly in $\theta \in [0, 1]$.

Since

$$\frac{\partial r_u}{\partial t_i}$$

is continuous on the compact cube $[0, 1]^k$, it is bounded and uniformly continuous. Therefore,

$$\frac{\partial r_u}{\partial t_i}\left(Z^{(a)}(q) + \frac{\theta}{m_i} e_i\right) \xrightarrow{P} \frac{\partial r_u}{\partial t_i}(q\mathbf{1}). \tag{15}$$

Since the derivative is bounded and uniformly continuous, Eq. (15) together with convergence in probability implies convergence in L^1 of the bounded integrand. Hence its conditional expectation converges to $\partial_i r_u(q\mathbf{1})$. The resulting quantities are uniformly bounded, so dominated convergence may then be applied to the outer integration over q. We obtain

$$\lim_{\min_{j \in u} m_j \to \infty} S_{i \to u}^{\mathrm{AS}}(\mathbf{m}) = \int_0^1 \frac{\partial r_u}{\partial t_i}(q\mathbf{1})\, dq. \tag{16}$$

Renaming q as τ gives

$$\lim_{\min_{j \in u} m_j \to \infty} S_{i \to u}^{\mathrm{AS}}(\mathbf{m}) = \int_0^1 \frac{\partial r_u}{\partial t_i}(\tau\mathbf{1})\, d\tau.$$

It remains to establish efficiency of the limit. Define

$$h(\tau) := r_u(\tau\mathbf{1}).$$

By the chain rule,

$$h'(\tau) = \sum_{i \in u} \frac{\partial r_u}{\partial t_i}(\tau\mathbf{1}).$$

Therefore

$$\begin{aligned}\sum_{i \in u} \int_0^1 \frac{\partial r_u}{\partial t_i}(\tau\mathbf{1})\, d\tau &= \int_0^1 h'(\tau)\, d\tau \\ &= r_u(\mathbf{1}) - r_u(\mathbf{0}).\end{aligned} \tag{17}$$

By boundary vanishing,

$$r_u(\mathbf{0}) = 0,$$

and by endpoint recovery,

$$r_u(\mathbf{1}) = \phi_u.$$

Hence

$$\sum_{i \in u} \int_0^1 \frac{\partial r_u}{\partial t_i}(\tau\mathbf{1})\, d\tau = \phi_u.$$

□

B.4 Proof of Proposition 3

Let

$$n := |N'_u| = \sum_{j \in u} m_j.$$

Fix a feature $i \in u$ and one of its micro-players

$$a = (i, s).$$

The standard Shapley formula gives

$$\varphi_a^{\mathrm{Sh}}(v_u) = \sum_{A \subseteq N'_u \setminus \{a\}} \frac{|A|!\,(n - |A| - 1)!}{n!} \times \left[v_u(A \cup \{a\}) - v_u(A) \right].$$

By the definition of the induced game,

$$v_u(A) = r_{p(A)}.$$

If

$$p(A) = p,$$

then

$$|A| = |p|$$

and, because a belongs to feature i,

$$v_u(A \cup \{a\}) - v_u(A) = r_{p+e_i} - r_p = \Delta_p(i).$$

We may therefore group all coalitions having the same grid state.

For a fixed state

$$p \in \mathcal{G}_{\mathbf{m}}(u) \qquad \text{with} \qquad p_i < m_i,$$

the number of coalitions

$$A \subseteq N'_u \setminus \{a\}$$

satisfying $p(A) = p$ is

$$\binom{m_i - 1}{p_i} \prod_{\substack{j \in u \\ j \neq i}} \binom{m_j}{p_j}. \tag{18}$$

Consequently,

$$\varphi_a^{\mathrm{Sh}}(v_u) = \sum_{p \in \mathcal{G}_{\mathbf{m}}(u) p_i < m_i} \frac{|p|!\,(n - |p| - 1)!}{n!} \binom{m_i - 1}{p_i} \times \left(\prod_{\substack{j \in u \\ j \neq i}} \binom{m_j}{p_j} \right) \Delta_p(i).$$

The m_i micro-players belonging to feature i are symmetric. Summing Eq. (18) over $s = 1, \ldots, m_i$ therefore multiplies the expression by m_i:

$$S_{i \to u}^{\mathrm{AS}}(\mathbf{m}) = \sum_{\substack{p \in \mathcal{G}_{\mathbf{m}}(u) \\ p_i < m_i}} \frac{|p|!\,(n - |p| - 1)!}{n!} m_i \binom{m_i - 1}{p_i} \times \left(\prod_{\substack{j \in u \\ j \neq i}} \binom{m_j}{p_j} \right) \Delta_p(i).$$

Finally, use the binomial identity

$$m_i \binom{m_i - 1}{p_i} = (m_i - p_i) \binom{m_i}{p_i}. \tag{19}$$

Substitution into Eq. (18) gives

$$S^{\mathrm{AS}}_{i\to u}(\mathbf{m}) = \sum_{\substack{p\in\mathcal{G}_{\mathbf{m}}(u)\\ p_i<m_i}} \frac{|p|!\,(n-|p|-1)!}{n!} \times \left(\prod_{j\in u}\binom{m_j}{p_j}\right)(m_i-p_i)\Delta_p(i),$$

which is exactly Eq. (11).

In the uniform-resolution case $m_j = m$ for all $j \in u$, the number of grid states is

$$|\mathcal{G}_{\mathbf{m}}(u)| = (m+1)^k.$$

For every state we evaluate at most k feature directions, so once the residual grid has been constructed, the exact attribution step costs

$$\mathcal{O}\big(k(m+1)^k\big)\,.$$

By contrast, direct enumeration operates on

$$n = km$$

micro-players and costs

$$\mathcal{O}(km\,2^{km}).$$

□

C Generalization to LES Values

Aumann-SHAP uses the Shapley value to allocate the induced micro-game v_u, but the geometric construction itself is not specific to Shapley. The same interaction surface and micro-player game can be combined with any Linear–Efficient–Symmetric (LES) value. This provides alternative within-pot allocation rules while leaving the interaction pot ϕ_u unchanged.

Let $n := |N'_u|$. An LES value can be written as

$$\varphi^{\mathrm{LES}}_a(v_u) = \sum_{A\subseteq N'_u\setminus\{a\}} \frac{|A|!\,(n-|A|-1)!}{n!} \times \Big[b(|A|+1)v_u(A\cup\{a\}) - b(|A|)v_u(A)\Big], \tag{20}$$

where $a = (i,s)$, $b(0) = 0$, $b(n) = 1$, and the intermediate coefficients $b(1),\dots,b(n-1)$ determine the particular LES rule (Ruiz et al., 1998; Chameni Nembua & Andjiga, 2008; Chameni Nembua, 2012; Radzik & Driessen, 2013; Béal et al., 2015).

As for Aumann-SHAP, micro-player payoffs are aggregated back to the original feature:

$$S^{\mathrm{LES}}_{i\to u}(\mathbf{m}) := \sum_{s=1}^{m_i} \varphi^{\mathrm{LES}}_{(i,s)}(v_u).$$

Efficiency gives $\sum_{i\in u} S^{\mathrm{LES}}_{i\to u}(\mathbf{m}) = \phi_u$. Thus changing the LES rule changes only how the fixed pot is divided, not the quantity being explained.

The Shapley case used throughout the main theory corresponds to $b(s) = 1$ for $1 \le s \le n-1$. Other choices recover alternative symmetric efficient rules; in particular, Equal Surplus is included as an auxiliary comparison in the MNIST experiment. We leave systematic comparison of LES refinements to future work, since the focus of the paper is the discrete-to-continuous Shapley bridge.

D Additional Experimental Details

Analytical interaction. The synthetic experiment uses $g(x_1,x_2,x_3) = x_1x_2^2 + x_3$ with $\mathbf{x}^0 = (0,0,0)$ and $\mathbf{x}^1 = (1,1,1)$. For $u = \{1,2\}$, the pure interaction surface is $r_u(t_1,t_2) = t_1t_2^2$, so $\phi_u = 1$. The diagonal Aumann–Shapley shares are analytically $(1/3, 2/3)$. Finite-resolution Aumann-SHAP is evaluated under uniform refinement and compared with the $m = 1$ equal split; the reported curve reaches approximately $(0.333, 0.667)$ at $m = 50$.

UCI Heart Disease. We use the Cleveland Heart Disease data with 303 observations and 13 features. The fitted XGBoost model achieves AUC 0.91. The reported transition involves $u = \{\texttt{chol}, \texttt{thalach}, \texttt{oldpeak}\}$ and moves from $(240, 171, 0.9)$ to $(258, 157, 2.6)$. The corresponding interaction pot is $\phi_u = 0.0103$. All reported values are model attributions for this fixed transition; they are not interpreted as causal or clinical effects.

German Credit. The German Credit data contain 1,000 applicants and 27 attributes. For the representative Growing-Spheres transition, the changed-feature set is $\{X_8, X_{13}, X_{14}, X_{25}\}$, the prediction change is $\Delta V = 0.6569$, and Aumann-SHAP uses uniform resolution $m = 5$. For the pairwise pot $\{X_8, X_{13}\}$, equal splitting gives $(0.0659, 0.0659)$, whereas Aumann-SHAP gives $(0.1143, 0.0176)$.

For the dataset-level analysis, we consider 138 low-score XGBoost baselines satisfying $p_{\text{xgboost}} < 0.30$ for which a successful counterfactual with $p_{\text{xgboost}}(\mathbf{x}^1) \geq 0.80$ is obtained. The successful transitions change 2.51 features on average. We report the fraction for which the strict feature-priority ordering under Aumann-SHAP differs from the equal-split ordering: $17/138 = 12.3\%$.

MNIST. MNIST contains 60,000 training and 10,000 test images. A ResNet-18 classifier achieves test accuracy 0.9972. For the representative $1 \to 7$ transition used in the patch test, we retain pixels satisfying $|x^1_{rc} - x^0_{rc}| > 0.05$, yielding 109 editable pixels. Aumann-SHAP uses $m = 10$; because the corresponding micro-game is too large for exact enumeration, Shapley values are estimated using 200 random permutations.

For the global visualization, we sample 200 test-set images of digit 1 and pair each baseline with its Euclidean nearest neighbor among test-set images of digit 7. Attributions are computed on the pair-specific changed-pixel set using the same threshold 0.05, and the resulting attribution maps are averaged over the 200 pairs.

In the patch test, all methods receive exactly the same editable pixels. They differ only in the order in which target pixel values are inserted into the baseline image. After the top-K edits are applied, we record the target probability P_7. The reported AUC is the area under this probability-versus-budget curve. Random ordering is averaged over 200 trials.

Runtime comparison. The runtime experiment compares direct enumeration of the micro-coalition game with the exact grid-state formulation. Direct enumeration operates on $n = km$ micro-players and scales as $\mathcal{O}(km\, 2^{km})$, whereas the grid-state implementation enumerates $(m+1)^k$ states and evaluates at most k directions per state, giving $\mathcal{O}(k(m+1)^k)$. The runtime plots vary resolution m and interaction order k separately to illustrate these two scaling regimes.