SHAPLEY IDENTIFICATION UNDER PARTIAL INFORMATION IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In privacy-preserving federated learning, clients can evaluate two utilities locally without revealing their updates: the utility of their update alone and the utility lost when they are removed from the full aggregate. A scorer can collect these measurements across clients. We study how much this collection determines the full vector of Shapley values and develop a theory of what can be inferred from these measurements. In one completed federated round, their midpoint recovers Shapley whenever interactions involve at most two clients. With three-client interactions, however, the same observations can correspond to different Shapley values. For games with nonnegative effects involving at most three clients, we characterize each client's interval of compatible Shapley values and the rankings shared by all compatible Shapley vectors. Beyond the pairwise case, smoothness still controls the error: the midpoint differs from the true Shapley value only at cubic order in the update size, and no rule using the same measurements can improve this order. Given bounds on curvature and update size, this approximation yields certified score intervals. In our experiments, these intervals contain every exact Shapley value and determine nearly all pairwise client rankings.

1 INTRODUCTION

Cross-silo federated learning (FL) consortia often need to attribute the improvement of a shared model to individual clients, whether to divide its benefits or to identify the clients behind an update that degraded it. The MELLODDY project (Heyndrickx et al., 2024) illustrates why benefit distribution matters: ten pharmaceutical companies privately trained drug-discovery models and all benefited, though unequally. The authors warned that granting all partners equal access to the resulting models despite differences in their contributions to training may discourage further data sharing.

Cooperative game theory formalizes this attribution problem by giving every subset of clients a *worth*, the utility gain its updates produce, and scoring each client from those worths. The Shapley value averages a client's marginal contribution over all orderings of the clients (Shapley, 1953), and it is the unique score that is efficient, anonymous, linear, and zero for a client that never changes the utility. This makes it the standard rule for contribution scoring in FL and for data valuation (Ghorbani & Zou, 2019; Zheng et al., 2023).

Computing the Shapley value requires the worth of many subsets, and which of those worths are available depends on how the round is aggregated and on what the scoring protocol authorizes. In standard FL the server observes every individual update and its aggregation weight, so it can assemble the model of any subset of clients and exact scoring is only a matter of cost, commonly addressed by sampling subsets. Under secure aggregation the server observes only the combined update (Bonawitz et al., 2017) and can form no subset model at all. Privacy-preserving scoring protocols sit between these extremes, keeping individual updates hidden while making a chosen family of coalition evaluations available (Zheng et al., 2023). Once only some coalitions can be evaluated, the prior question is whether the evaluations a protocol provides determine Shapley at all.

In this work, we study the measurement policy behind existing private-FL contribution scores (Pejó et al., 2021; Jaramillo Velez et al., 2026): each client evaluates the gain produced by its own update alone, and the loss when that update is removed from the full aggregate. *Do these two measurements per client determine the Shapley value?* We analyze this question under the most favorable access. We give the scorer both measurements for every client and the worth of the grand coalition, which

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.

together form the *endpoint observation*. If these pooled measurements do not identify Shapley, neither can a protocol limited to them.

The answer turns on interaction order. A subset's worth decomposes into effects: what each client achieves alone, what each pair adds beyond its members, what each triple adds beyond its pairs, and so on. Clients *interact* when the group they form adds something beyond its subgroups. Shapley splits each effect among the clients producing it, in halves for a pair and thirds for a triple, while the two evaluations reveal only a client's own effect and the undivided total of those it shares. The midpoint of a client's two evaluations, used as Fair-Private's unnormalized score (Jaramillo Velez et al., 2026), assigns the client its full individual effect and half of every shared effect: exactly the share Shapley assigns within a pair, and more than the third it assigns within a triple.

Our contribution. We show that the endpoint observation determines Shapley for pairwise interactions, characterize the exact identified set when effects are nonnegative and involve at most three clients, and derive certified approximation intervals under smoothness. Specifically:

- We prove that the endpoint observation determines Shapley whenever clients interact only in pairs, for any number of clients, and whenever a round has at most three clients (Cor. 4). For four or more clients, we prove that Shapley is not always determined once interactions among three clients are allowed (Prop. 5, Fig. 1).
- When effects are beneficial and involve at most three clients, the endpoint observation still determines the total pair and triple effects, though not which clients produced them. We derive the resulting *identified set*, the Shapley vectors of every game consistent with the observation, in closed form: a sharp interval for every client, and the rankings that hold across the set (Thm. 6).
- We characterize the scores computable from the endpoint observation. Linearity, equal treatment, and correct credit for null and constant contributors force each client's score to use only its own two measurements, and symmetry between them selects the midpoint (Prop. 11).
- We bound the midpoint's error for smooth utilities by a cubic term in update size, the best order attainable from these measurements, and turn that bound into certified per-client intervals. In prespecified MHEALTH rounds they cover exact Shapley in all 20 games, certify over 99% of signs and rankings, and identify every top client (Thms. 7 and 8).

2 Related work

Contribution scoring in federated learning. The Shapley value is the standard basis for data valuation (Ghorbani & Zou, 2019) and for scoring FL clients. A server holding the individual updates can evaluate any coalition, at a cost that sampling keeps manageable (Okhrati & Lipani, 2021); secure aggregation removes that access (Bonawitz et al., 2017). Private protocols therefore authorize a restricted family of evaluations: Zheng et al. (2023) compute Shapley securely from many of them, while Pejó et al. (2021) use two per client, the gain from a client's update alone and the loss when it is removed, whose midpoint Jaramillo Velez et al. (2026) normalize into the Fair-Private score.

Values from partial information. Cooperative game theory studies values for games known on part of the coalition lattice, ranges over their completions, and extensions respecting positivity (Willson, 1993; Housman, 2002; Masuya & Inuiguchi, 2016; Masuya, 2021; Černý & Grabisch, 2024; Černý et al., 2024; Černý, 2025); Bok & Černý (2024, Section 5) already permits our coalitions under 1-convexity. Here the observed coalitions come from an FL protocol and our axioms constrain values on complete games (App. D); semivalues contain the two endpoint rules (Dubey et al., 1981; Fried, 2026).

Interaction order. Low-order surrogates keep only the leading interaction terms (Sundararajan et al., 2020; Tsai et al., 2023; Pelegrina et al., 2026; Xu et al., 2025). We instead tie interaction order to a fixed measurement policy: order two is where these two measurements stop determining Shapley.

3 Preliminaries and Problem Setup

Federated round and scoring. We attribute the utility change of one completed cross-silo FL round with clients $\mathcal{N} = \{1, \dots, n\}$. Let M_t be the global model at the start of round t. Client i uses

its dataset D_i to compute an update $U_{i,t}$ from M_t and the server combines these updates with public nonnegative weights q_i into the next global model,

$$M_{t+1} = M_t + \sum_{i \in \mathcal{N}} q_i U_{i,t} \,.$$

For data-proportional Federated Averaging (FedAvg), $q_i = |D_i| / \sum_j |D_j|$.

We measure model quality with a common utility function F_t, such as negative validation loss, where larger values are better. Secure aggregation reveals M_{t+1} while hiding the individual updates (Bonawitz et al., 2017). The utility evaluations used for scoring must be supplied separately. We assume clients carry out these evaluations honestly.

Fixed-update contribution game. For a subset $\mathcal{S} \subseteq \mathcal{N}$ of clients, called a *coalition*, the model produced by its updates alone is

$$M_t(\mathcal{S}) = M_t + \sum_{i \in \mathcal{S}} q_i U_{i,t} \,.$$

At the extremes, $M_t(\varnothing) = M_t$ is the starting model and $M_t(\mathcal{N}) = M_{t+1}$ is the released one. The *worth* of $\mathcal{S}$ is the utility this model gains over M_t,

$$v_t(\mathcal{S}) = F_t(M_t(\mathcal{S})) - F_t(M_t) \,, \tag{1}$$

so $v_t(\varnothing) = 0$. We call $\mathcal{N}$ the *grand coalition* and $V_t = v_t(\mathcal{N})$ the utility gain of the completed round.

Our aim is to attribute the round's utility change to its clients, including any negative contributions.

Shapley attribution. From here on, we suppress the round index and write v for the contribution game, and we let $\Gamma^{\mathcal{N}}$ denote the games $v : 2^{\mathcal{N}} \to \mathbb{R}$ with $v(\varnothing) = 0$. Fix an ordering of the clients: the marginal worth of client i is what it adds when it joins those before it. The Shapley value $\phi_i(v)$ averages the marginal worth of client i over all orderings (Shapley, 1953),

$$\phi_i(v) = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{S}|!\,(n - |\mathcal{S}| - 1)!}{n!} \big(v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})\big) \,. \tag{2}$$

We write $\Psi : \Gamma^{\mathcal{N}} \to \mathbb{R}^n$ for a generic contribution value, and reserve *worth* for the scalar $v(\mathcal{S})$.

Endpoint observation. We study a scoring policy that provides the scorer with two measurements per client, as in existing private-FL contribution methods (Pejó et al., 2021; Jaramillo Velez et al., 2026),

$$s_i = v(\{i\}), \qquad m_i = v(\mathcal{N}) - v(\mathcal{N} \setminus \{i\}) \,, \tag{3}$$

the client's *standalone gain* and *removal loss*. Client i can form both models from quantities it already holds, since $M_t(\{i\}) = M_t + q_i U_{i,t}$ and $M_t(\mathcal{N} \setminus \{i\}) = M_{t+1} - q_i U_{i,t}$, so obtaining them reveals no other client's individual update. These are the two ends of the joining order in (2): s_i is the marginal worth of client i when it joins the empty coalition, and m_i its marginal worth when it joins all the others. Together with the grand-coalition worth $V = v(\mathcal{N})$ they form the *endpoint observation*

$$O_{\mathrm{end}}(v) = (V, (s_i)_{i \in \mathcal{N}}, (m_i)_{i \in \mathcal{N}}) \,. \tag{4}$$

We give the scorer access to the full observation $O_{\mathrm{end}}(v)$.

Definition 1. A value $\Psi : \Gamma^{\mathcal{N}} \to \mathbb{R}^n$ is *endpoint-local* if some map ψ satisfies $\Psi(v) = \psi(O_{\mathrm{end}}(v))$ for every game v, and *client-local* if the score of client i depends only on its own pair (s_i, m_i).

An endpoint-local score may therefore combine every client's measurements. The scoring procedure can hold the observation directly or work with it inside a protected computation, so access to it need not mean disclosure to the server or to the other clients.

For our impossibility results, we give the scorer the complete endpoint observation together with the public round information: the model M_t, the utility F_t, the weights $(q_i)_i$, and the released aggregate M_{t+1}.

4 WHAT THE ENDPOINT OBSERVATION DETERMINES ABOUT SHAPLEY

In this section, we determine when the endpoint observation suffices to recover Shapley. We say that Shapley is *identified* within a class of games when any two games in that class with the same observation have the same Shapley vector, and call such games *compatible*. We show that identification holds when interactions involve at most two clients (Sec. 4.1), but fails from four clients on when three-client interactions are allowed, even if every effect is beneficial and the public round information is fixed (Sec. 4.2). For positive 3-additive games, we then characterize every Shapley vector compatible with the observation (Sec. 4.3).

4.1 SHAPLEY IS IDENTIFIED FOR PAIRWISE INTERACTIONS

To compare the Shapley value with the two measurements, we write both as weighted sums of the same quantities, the *Harsanyi dividends* (Harsanyi, 1963), which decompose worth into interaction effects. The dividend $d_v(\mathcal{T})$ of a coalition $\mathcal{T}$ is the worth left after subtracting the effects of all its proper subsets, so a singleton dividend is a client's standalone worth and a pair dividend is what the pair adds beyond its two members. Dividends reconstruct every worth, and the Shapley value splits each of them equally among the clients producing it,

$$v(\mathcal{S}) = \sum_{\mathcal{T}\subseteq\mathcal{S}} d_v(\mathcal{T}), \qquad d_v(\mathcal{T}) = \sum_{\mathcal{S}\subseteq\mathcal{T}} (-1)^{|\mathcal{T}|-|\mathcal{S}|} v(\mathcal{S}), \qquad \phi_i(v) = \sum_{\mathcal{T}\ni i} \frac{d_v(\mathcal{T})}{|\mathcal{T}|}\,.$$

We call v *k-additive* when $d_v(\mathcal{T}) = 0$ for every $|\mathcal{T}| > k$, so no group larger than k has an effect beyond those of its subgroups (Grabisch, 1997). For example, a pair whose joint effect is worth 6 adds 3 to each of its two clients' scores, while a triple whose joint effect is worth 6 adds 2 to each of its three. App. A.3 shows that this dividend form of ϕ agrees with the ordering average (2).

The two measurements in (3) are sums of the same dividends, weighted differently: the standalone gain of client i is its own effect, and removing i removes every effect it takes part in, so

$$s_i = d_v(\{i\}), \qquad m_i = \sum_{\mathcal{T}\ni i} d_v(\mathcal{T})\,.$$

The Shapley value assigns client i half of every pair effect it takes part in, which is exactly what halving $m_i - s_i$ yields. The next proposition states the resulting recovery: with only pairwise interactions, the midpoint is the Shapley value.

Proposition 2. *Let v be 2-additive. Then, for every number of clients and every $i \in \mathcal{N}$,*

$$\phi_i(v) = s_i + \frac{1}{2}(m_i - s_i) = \frac{s_i + m_i}{2} = \Psi_i^{1/2}(v)\,, \tag{5}$$

the midpoint *of the two measurements.*

The proof is a two-line dividend computation (App. A.3). Thus the midpoint is client-local: each score uses only (s_i, m_i). An effect involving $k \geq 3$ clients breaks this match: the midpoint still assigns each of them half of it, whereas the Shapley value assigns a $1/k$ share. Summing these discrepancies gives the exact error of the midpoint,

$$\Psi_i^{1/2}(v) - \phi_i(v) = \sum_{\substack{\mathcal{T}\ni i \\ |\mathcal{T}|\geq 3}} \left(\frac{1}{2} - \frac{1}{|\mathcal{T}|}\right) d_v(\mathcal{T})\,. \tag{6}$$

Every term of (6) involves three or more clients, so the midpoint is exact on 2-additive games and departs from Shapley only through effects of order three and above. This describes the midpoint. A different rule, built from V and every client's measurements rather than from one client's pair, might still recover Shapley; we rule this out next.

4.2 WHY TRIPLE EFFECTS CAN HIDE SHAPLEY

For four or more clients, we show that the endpoint observation in (4) need not determine Shapley when three-client interactions are allowed. This remains true even when the utility, weights, and released aggregate are held fixed.

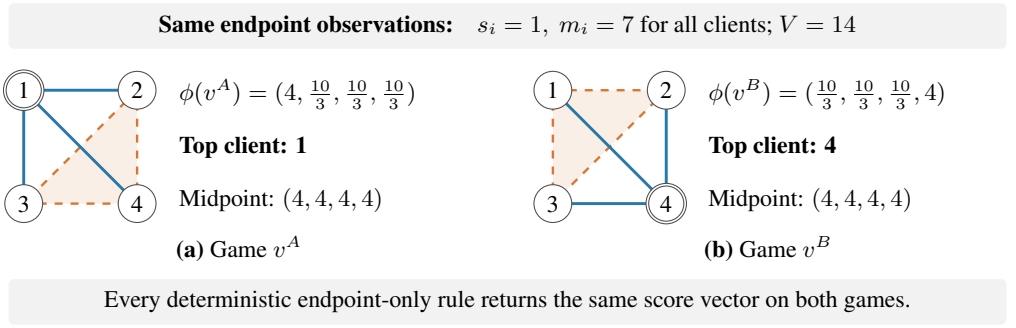


Figure 1: Two four-client games with the same endpoint observation but different Shapley values and top-ranked clients.

We call a game *positive* when every effect is nonnegative (Černý et al., 2024). In a positive 3-additive game, then, every effect is beneficial and no group of four or more clients has an effect of its own.

Example. Figure 1 shows two four-client games with identical observations: in both, every client has standalone worth 1, removing any one client costs the grand coalition 7, and that coalition is worth 14. The games place the same total effect at each client, and differ only in how that total splits between pairs and triples. Since the Shapley value weights a pair effect by one half and a triple effect by one third, this split changes the scores: client 1 ranks first on the left and client 4 on the right. Thus, no rule based only on the endpoint observation can recover Shapley in both games or identify the top-ranked client.

The following theorem shows that this ambiguity is not confined to four clients.

Theorem 3. *For every $n \geq 4$, there are positive 3-additive games v, w such that $O_{\text{end}}(v) = O_{\text{end}}(w)$ and $\phi(v) \neq \phi(w)$.*

The proof is given in App. A.4.

Corollary 4. *For at most three clients, the endpoint observation identifies Shapley without any restriction on interaction order. For $n \geq 4$ and $1 \leq k \leq n$, it identifies Shapley within the class of k-additive games if and only if $k \leq 2$.*

For at most three clients, the endpoint observation $O_{\text{end}}(v)$ contains every nonempty coalition worth: s_i gives singletons, $V - m_i$ their complements, and V the grand coalition. These are all the nonempty coalition worths, so the observation determines Shapley directly. For $n \geq 4$, Prop. 2 guarantees recovery for pairwise interactions, but not over the full k-additive class when $k \geq 3$: the games in Thm. 3 share $O_{\text{end}}(v)$ but have different Shapley values.

We now give an FL realization of Thm. 3. We call the collection $(U_i)_{i \in \mathcal{N}}$ of client updates in one round an *update profile*. The next proposition constructs two such profiles with the same starting model, utility, weights, and released aggregate.

Proposition 5. *Fix positive public weights (q_i). There are two update profiles $U = (U_i)_{i \in \mathcal{N}}$ and $U' = (U'_i)_{i \in \mathcal{N}}$ and one degree-three polynomial utility F such that, from the same starting model $\theta = 0$, the profiles induce the games of Thm. 3 and satisfy*

$$\sum_i q_i U_i = \sum_i q_i U'_i \,.$$

The proof is given in App. A.4.1. Under the two profiles of Prop. 5, the scorer sees the same starting model, utility, weights, released aggregate, and endpoint observation, but the true Shapley vectors differ. A deterministic rule based on these inputs must therefore return the same estimate for both and cannot be exact on both. Cor. 16 in App. A.4.1 quantifies this error: in any norm, at least one profile has error at least half the distance between the two true Shapley vectors. Sec. 5.3 shows that identical information available to the scorer can still correspond to different Shapley vectors when every client update is multiplied by a common factor $\varepsilon \to 0$, yielding a lower bound proportional to ε^3 on endpoint-based estimation error.

4.3 SHAPLEY INTERVALS FOR POSITIVE 3-ADDITIVE GAMES

Sec. 4.2 shows that one endpoint observation can be compatible with different Shapley vectors. Here we characterize the *identified set*, the vectors compatible with that observation, for positive 3-additive games. Such games have only standalone, pair, and triple effects, all nonnegative. To isolate what clients achieve together, subtract the observed standalone effects,

$$J = V - \sum_i s_i, \qquad r_i = m_i - s_i, \qquad D = \sum_i r_i \,.$$

Let P and H be the total pair and triple effects. Then $J = P + H$, while $D = 2P + 3H$: each r_i counts the shared effects involving i, so their sum counts every pair twice and every triple three times. Thus the observation determines both totals,

$$P = 3J - D, \qquad H = D - 2J \,.$$

The observation fixes each r_i as well, but not its split between pair and triple effects. Let t_i be the total triple effect involving client i. Its total pair effect is $r_i - t_i$, so Shapley assigns

$$\phi_i = s_i + \frac{r_i - t_i}{2} + \frac{t_i}{3} = \Psi_i^{1/2} - \frac{t_i}{6} \,.$$

Nonnegativity requires $0 \le t_i \le H$ and $0 \le r_i - t_i \le P$; counting each triple at its three clients gives $\sum_i t_i = 3H$. The theorem below shows which such allocations can arise from a positive 3-additive game and which Shapley vectors they yield. (The proof is given in App. A.5)

Theorem 6. *For $n \ge 3$, let $o = (V, (s_i)_i, (m_i)_i)$ be an endpoint observation and define*

$$\ell_i = \max\{0, r_i - P\}, \qquad u_i = \min\{H, r_i\} \,.$$

The endpoint observation has a positive 3-additive extension if and only if $s_i \ge 0$, $P, H \ge 0$, $\ell_i \le u_i$ for every i, and

$$\sum_i \ell_i \le 3H \le \sum_i u_i \,.$$

When feasible, the complete Shapley identified set is

$$\mathcal{I}(o) = \Big\{ \Psi^{1/2}(o) - \frac{t}{6} : \ell_i \le t_i \le u_i, \quad \sum_i t_i = 3H \Big\}, \qquad \Psi_i^{1/2}(o) = (s_i + m_i)/2 \,. \tag{7}$$

Projecting the identified set in Thm. 6 onto coordinate i gives the per-client bounds. Let $t_i^- = \max\{\ell_i, 3H - \sum_{j \ne i} u_j\}$ and $t_i^+ = \min\{u_i, 3H - \sum_{j \ne i} \ell_j\}$. Then every compatible positive 3-additive game satisfies

$$s_i + \frac{r_i}{2} - \frac{t_i^+}{6} \le \phi_i \le s_i + \frac{r_i}{2} - \frac{t_i^-}{6} \,.$$

Both endpoints are attained, so the bounds are tight.

Since $\sum_j t_j = 3H$, t_i^- is the smallest total triple effect involving client i consistent with the other clients' upper bounds, and t_i^+ is the largest consistent with their lower bounds. The per-client intervals do not vary independently: a Shapley vector is compatible exactly when each coordinate lies in its interval and the coordinates sum to V. For client intervals $[a_i, b_i]$ and $[a_j, b_j]$, the exact range of $\phi_i - \phi_j$ is $[a_i - b_j, b_i - a_j]$. Thus i strictly outranks j in every compatible game exactly when $a_i > b_j$ (App. A.5.1).

Since $t_i \ge 0$, the midpoint bounds Shapley from above in this class. When $H = 0$, every interval collapses to the pairwise value. Computing the bounds requires the complete endpoint observation. To apply these bounds to an FL game, we must separately establish that every effect is nonnegative and involves at most three clients (App. C.3).

As we show in App. A.4, for $n \ge 4$, the Shapley vectors compatible with any fixed endpoint observation are unbounded when interaction effects are unrestricted in sign and magnitude.

Fig. 2 illustrates Thm. 6 for four clients with $s_i = 1$ and $m_i = 7$ fixed, showing how each client's compatible Shapley interval changes with the observed grand-coalition worth V. The shaded band gives the exact interval over positive 3-additive games, while the dashed line marks the midpoint

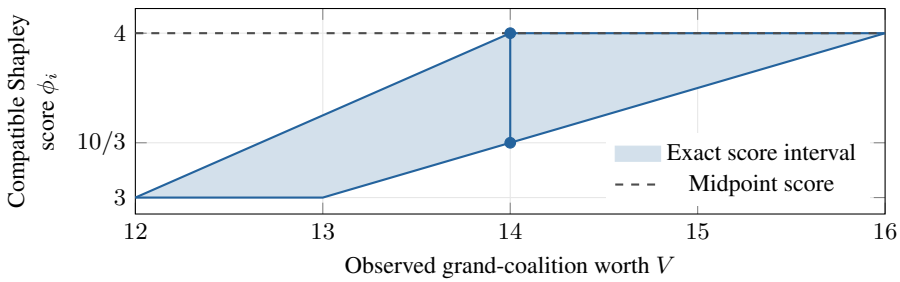


Figure 2: Compatible Shapley intervals as the grand-coalition worth varies.

score 4. The midpoint remains 4 as V varies, but V changes P, H, and therefore the exact compatible interval. At $V = 14$, as in Fig. 1, $P = 6$, $H = 4$, and the interval is $[10/3, 4]$. Although 4 lies in every client's interval, all four Shapley values cannot equal 4 because they must sum to $V = 14$. Thus the full endpoint observation matters, and the coordinate intervals are coupled by efficiency. At $V = 16$, $H = 0$, and every interval collapses to the pairwise midpoint.

5 Approximating Shapley from the Observation

Sec. 4 shows when the endpoint observation determines Shapley and characterizes the compatible Shapley values for positive 3-additive games. We now ask how accurately a rule based on this observation can approximate Shapley when it is not determined. For $n \geq 4$, we show that attribution axioms restrict each score to a linear combination of the client's two measurements. Within this family, only the midpoint is exact for every pairwise-interaction game, and it treats the measurements symmetrically (Sec. 5.1). For smooth utilities, given bounds on curvature and update size, we derive a cubic upper bound on the midpoint's absolute error (Sec. 5.2). We then prove that no rule given the endpoint observation and all public round information can guarantee a better worst-case error order (Sec. 5.3).

5.1 From attribution axioms to the midpoint

We first characterize endpoint-local scores that satisfy four attribution axioms, then use pairwise exactness to identify the midpoint among them. The axioms we enforce are linearity, anonymity (equal treatment under client renaming), the null-player axiom (zero credit when a client never changes utility), and dummy calibration (credit x when a client's marginal contribution is always x). For $n \geq 4$, Prop. 11 in App. A.2 shows that these requirements restrict every endpoint-local rule to

$$\Psi_i^{(c)} = (1 - c)s_i + c \cdot m_i, \qquad c \in \mathbb{R},$$

where $s_i = v(\{i\})$ is the gain from client i alone and $m_i = V - v(\mathcal{N} \setminus \{i\})$ is the loss from removing it from the full group. Thus the four axioms make every endpoint-local rule client-local, despite access to the complete endpoint observation. The resulting scores need not sum to V.

On pairwise-interaction games, $\phi_i = s_i + \frac{1}{2}(m_i - s_i)$ (Prop. 2). Since $m_i - s_i$ can be nonzero, the midpoint $\Psi_i^{1/2} = (s_i + m_i)/2$ is the only member of this family that is exact for all such games.

The midpoint is also *self-dual*: its scores are unchanged when v is replaced by the dual game $v^D(\mathcal{S}) = V - v(\mathcal{N} \setminus \mathcal{S})$, which exchanges s_i and m_i. We prove in Cor. 12 (App. A.2) that it is the unique self-dual member of this family.

5.2 Cubic error and certified intervals

To bound the midpoint's error beyond pairwise interactions, let S_t include each client other than i independently with probability t, and define

$$f_i(t) = \mathbb{E}\big[v(S_t \cup \{i\}) - v(S_t)\big].$$

Its endpoints are s_i and m_i. Integrating the inclusion probabilities gives the Shapley weights (Okhrati & Lipani, 2021), hence

$$\phi_i(v) = \int_0^1 f_i(t)\, dt. \tag{8}$$

The midpoint is the trapezoidal approximation to this integral. If $|f_i''(t)| \le C_i$ on $[0,1]$, the remainder places Shapley in

$$\phi_i(v) \in \left[\Psi_i^{1/2}(v) - \frac{C_i}{12}, \Psi_i^{1/2}(v) + \frac{C_i}{12}\right] . \tag{9}$$

This interval bound applies to signed interactions of any order. We next bound C_i using bounds on the utility's curvature and the update sizes.

Fix the checkpoint θ, utility $F : \mathbb{R}^d \to \mathbb{R}$, and weighted updates $z_i = q_i U_i$, with $z_S = \sum_{i\in\mathcal{S}} z_i$. Scaling all updates gives $v_\varepsilon(\mathcal{S}) = F(\theta + \varepsilon z_S) - F(\theta)$. Its linear Taylor term has only single-client effects, and its quadratic term has at most pairwise effects. Since the midpoint recovers both exactly (Prop. 2), its error begins at cubic order. We use a fixed update-space norm and its induced operator norms. We write $\widehat{\phi} = \Psi^{1/2}$ for the midpoint estimate and measure its relative error as $\|\widehat{\phi}-\phi\|_1/\|\phi\|_1$ when the denominator is nonzero.

Thm. 7 gives a client-specific error bound from curvature and update-size bounds and identifies the leading cubic term.

Theorem 7. *Suppose some compact convex K contains all segments from θ to $\theta + \varepsilon z_S$ for $\mathcal{S} \subseteq \mathcal{N}$ and $\varepsilon \in [0, \varepsilon_0]$, and F is C^3 near K. Let $L = \sup_K \|D^3 F\|_{\mathrm{op}}$, and suppose available bounds satisfy $a_i \ge \|z_i\|$. Then, throughout that range,*

$$|\Psi_i^{1/2}(v_\varepsilon) - \phi_i(v_\varepsilon)| \le \frac{L\varepsilon^3}{6}\, a_i \sum_{\substack{j<k \\ j,k\neq i}} a_j a_k .$$

Moreover,

$$\Psi_i^{1/2}(v_\varepsilon) - \phi_i(v_\varepsilon) = \frac{\varepsilon^3}{6} \sum_{\substack{j<k \\ j,k\neq i}} D^3F(\theta)[z_i, z_j, z_k] + o(\varepsilon^3) .$$

If $(DF(\theta)[z_i])_i \neq 0$, the relative ℓ_1 error is $O(\varepsilon^2)$.

App. B proves the theorem and shows that the leading error involves mixed third derivatives across three distinct clients. This explains the cubic absolute rate and, when the first-order Shapley signal is nonzero, the quadratic relative rate. If the linear term vanishes, the cubic absolute bound gives only $O(\varepsilon)$ relative error when the quadratic term leads, or $O(1)$ when the cubic term leads (App. B.2).

5.3 Optimality of the cubic rate

Thm. 7 gives a cubic upper bound. To test whether another rule can improve this order even with additional public information, we give the estimator the *augmented observation*: the complete endpoint observation together with the public checkpoint, utility, weights, and released aggregate. We then construct two fixed update profiles under one evaluator whose augmented observations agree at every common scale $0 < \varepsilon \le 1$, while their Shapley vectors differ at cubic order.

Theorem 8. *For every $n \ge 4$ and positive public weights (q_i), there exist a shared checkpoint, one degree-three polynomial utility, and two fixed update profiles inducing positive 3-additive games $v_\varepsilon, w_\varepsilon$. For every $0 < \varepsilon \le 1$, their augmented observations $x_\varepsilon = (O_{\mathrm{end}}(v_\varepsilon), \theta, F, (q_i), \varepsilon \sum_i q_i U_i)$ agree. Every estimator $\widehat{\phi}_\varepsilon$ based on this observation, optionally using independent randomness, satisfies*

$$\max_{g\in\{v_\varepsilon, w_\varepsilon\}} \mathbb{E}\|\widehat{\phi}_\varepsilon(x_\varepsilon) - \phi(g)\|_1 \ge \frac{\varepsilon^3}{12} .$$

Both profiles have nonzero first-order Shapley signal. For a constant $c_n > 0$ independent of the estimator and scale, their maximum expected relative ℓ_1 error is at least $c_n\varepsilon^2$. Expectations concern estimator randomness only; deterministic rules are included.

Because the two profiles give the same allowed inputs at every scale, an estimator has the same output distribution for both. Their Shapley separation forces an error of at least half that distance on one profile (App. B.1). With nonzero first-order Shapley signal, the lower bound also matches the midpoint's quadratic relative-error rate. In absolute error, the midpoint's cubic rate is worst-case optimal even with the additional public information in the augmented observation.

Table 1: Coverage, relative interval width, and certified decisions from midpoint intervals in 20 eight-client games per collection at $\varepsilon = 0.5$.

Collection	Coverage	Rel. width	Signs	Strict pairs	Top client
UCI-HAR development	20/20	2.62%	160/160	539/560	18/20
MHEALTH confirmation	20/20	1.24%	159/160	555/560	20/20

6 Midpoint accuracy and certified decisions in federated learning

We evaluate the midpoint in completed FL rounds on three datasets: Fashion-MNIST (Xiao et al., 2017), UCI-HAR (Reyes-Ortiz et al., 2013), and MHEALTH (Banos et al., 2014). Fashion-MNIST tests the small-update error rates predicted by Thm. 7. UCI-HAR and MHEALTH test whether intervals from the same theorem determine score signs, pairwise rankings, and the top-ranked client. We compute exact Shapley values by evaluating every coalition to assess both tests. The midpoint uses only the endpoint observation; its intervals also use curvature and update-size bounds.

We use negative multiclass Brier loss for Fashion-MNIST and negative cross-entropy for UCI-HAR and MHEALTH. Under either utility, higher worth means lower evaluation loss. We measure absolute error E_1 and relative error G_1,

$$E_1 = \|\widehat{\phi} - \phi\|_1, \qquad G_1 = \frac{E_1}{\|\phi\|_1 + 10^{-12}}.$$

Across 30 eight-client Fashion-MNIST profiles, we multiply each fixed set of client updates by a common scale ε. Thm. 7 gives a cubic absolute-error bound and, when the first-order Shapley signal is nonzero, a quadratic relative-error rate. Fig. 4 in App. C.2 shows median fitted slopes of 3.125 for E_1 and 2.110 for G_1 over $\varepsilon \leq 1/2$, consistent with those rates.

For the certificate study, we evaluate 20 eight-client rounds each on UCI-HAR and MHEALTH using multinomial logistic models. The protocol projects augmented features to unit norm, giving $L = 1/\sqrt{2}$, and clips local updates at 0.2 before applying the public weights, giving $a_i = 0.2q_i$. We fixed $\varepsilon = 0.5$ and the protocol on UCI-HAR, then applied them unchanged to MHEALTH. App. C.1 gives the model and assignment details.

With these bounds, Thm. 7 gives each client an interval around its midpoint score that contains the true Shapley value. The identity $\sum_i \phi_i = V$ tightens these intervals without affecting coverage. Table 1 reports coverage: the number of rounds in which all eight exact Shapley values lie in their respective intervals. It also reports median relative width, how many score signs and strict pairwise rankings the intervals determine, and how often they identify the top-ranked client. An interval excluding zero determines a score's sign. If one client's interval lies entirely above another's, their ranking is determined. If it lies above all other intervals, that client must rank first.

All 40 exact Shapley vectors fall inside the intervals, as Thm. 7 guarantees. Their median widths relative to $\|\phi\|_1$ are 2.62% on UCI-HAR and 1.24% on MHEALTH. They certify 160 of 160 and 159 of 160 signs, and 539 of 560 and 555 of 560 strict pairwise rankings, respectively. They identify the top client in 18 of 20 UCI-HAR rounds and all 20 MHEALTH rounds. Thus the Fashion-MNIST results support the local rates in Thm. 7, while the activity-recognition results show that its finite-scale bound is narrow enough to certify nearly all signs and pairwise rankings at $\varepsilon = 0.5$.

7 Conclusion

We prove that endpoint observations determine Shapley for pairwise interactions at any client count and for all rounds with at most three clients. For four or more clients, we construct beneficial triple effects that make values and rankings ambiguous and show that unconstrained effects make compatible scores unbounded. Under nonnegative effects involving at most three clients, we derive tight Shapley intervals and common rankings. For smooth FL utilities, we prove the midpoint's cubic worst-case error order is optimal and derive certified intervals from curvature and update-size bounds. Fashion-MNIST confirms the predicted scaling. Across 20 MHEALTH rounds, our intervals covered all exact Shapley vectors, certified 555 of 560 pairwise rankings, and identified every top client. **Limitations.** We focus on single-round attribution in cross-silo FL.

AI USE STATEMENT

Generative AI assisted code, verification, literature organization, and editing. The authors checked the resulting derivations, analyses, citations, and prose and take responsibility for the paper.

ETHICS STATEMENT

This work uses established public benchmark datasets. Contribution scores can affect compensation and institutional decisions, but attribution alone does not establish incentive compatibility. Using these scores as payments requires a separate mechanism, governance process, and privacy review.

REPRODUCIBILITY STATEMENT

The appendices give all assumptions and complete proofs. The supplementary repository (https://anonymous.4open.science/r/privacy-constrained-cooperative-values-2B0E) provides all the code to reproduce the results in the paper.

REFERENCES

Oresti Banos, Rafael Garcia, and Alejandro Saez. MHEALTH. UCI Machine Learning Repository, 2014.

Jan Bok and Martin Černý. 1-convex extensions of incomplete cooperative games and the average value. *Theory and Decision*, 96(2):239–268, 2024.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.

Martin Černý. Positivity of incomplete cooperative games revisited. *Operations Research Letters*, 60:107277, 2025.

Martin Černý and Michel Grabisch. Incomplete cooperative games with player-centered information. *Discrete Applied Mathematics*, 346:62–79, 2024.

Martin Černý, Jan Bok, David Hartman, and Milan Hladík. Positivity and convexity in incomplete cooperative games. *Annals of Operations Research*, 340(2–3):785–809, 2024.

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

Matthew Fried. What semivalues cannot see: The information content of anonymous marginal values, 2026. arXiv:2607.07013v1.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, 2019.

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

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

Wouter Heyndrickx, Lewis Mervin, Tobias Morawietz, et al. MELLODDY: Cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *Journal of Chemical Information and Modeling*, 64(7):2331–2344, 2024.

David Housman. Linear and symmetric allocation methods for partially defined cooperative games. *International Journal of Game Theory*, 30(3):377–404, 2002.

Delio Jaramillo Velez, Gergely Biczok, Alexandre Graell i Amat, Johan Ostman, and Balazs Pejo. Private and robust contribution evaluation in federated learning, 2026. arXiv:2602.21721v1.

Satoshi Masuya. An approximated shapley value for partially defined cooperative games. *Procedia Computer Science*, 192:100–108, 2021.

Satoshi Masuya and Masahiro Inuiguchi. A fundamental study for partially defined cooperative games. *Fuzzy Optimization and Decision Making*, 15(3):281–306, 2016.

Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate shapley values. In *International Conference on Pattern Recognition (ICPR)*, 2021.

Balázs Pejó, Gergely Biczók, and Gergely Ács. Measuring contributions in privacy-preserving federated learning. *ERCIM News*, (126):35, 2021.

Guilherme Dean Pelegrina, Patrick Kolpaczki, and Eyke Hüllermeier. Shapley value approximation based on k-additive games. In *AAAI Conference on Artificial Intelligence*, 2026.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human activity recognition using smartphones. UCI Machine Learning Repository, 2013.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning (ICML)*, 2020.

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

Stephen J. Willson. A value for partially defined cooperative games. *International Journal of Game Theory*, 21(4):371–384, 1993.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chu Xu, 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 (NeurIPS)*, 2025.

Shuyuan Zheng, Yang Cao, and Masatoshi Yoshikawa. Secure shapley value for cross-silo federated learning. *Proceedings of the VLDB Endowment*, 16(7):1657–1670, 2023.

A ENDPOINT IDENTIFICATION AND POSITIVE BOUNDS

This section supplies the complete proofs behind pairwise recovery, the first non-identification result, the endpoint-local representation, and the sharp positive third-order identified set.

A.1 OBSERVATION GEOMETRY AND IDENTIFICATION

How many independent quantities does the endpoint observation reveal? For at least three clients, each recorded worth concerns a different coalition. The following count also gives the number of ways a game can change without changing the observation.

Proposition 9 (Rank of the endpoint observation map). *For $n \geq 3$, $O_{\mathrm{end}} : \Gamma^{\mathcal{N}} \to \mathbb{R}^{2n+1}$ is surjective. Hence*

$$\operatorname{rank} O_{\mathrm{end}} = 2n+1, \qquad \dim \ker O_{\mathrm{end}} = 2^n - 2n - 2,$$

and $\Gamma^{\mathcal{N}} / \ker O_{\mathrm{end}} \simeq \mathbb{R}^{2n+1}$.

Proof of Proposition 9. For $n \geq 3$, the grand coalition, the n singletons, and the n leave-one-out coalitions are distinct. Given arbitrary $(V, (s_i), (m_i)) \in \mathbb{R}^{2n+1}$, set $v(\mathcal{N}) = V$, $v(\{i\}) = s_i$, and $v(\mathcal{N} \setminus \{i\}) = V - m_i$, and assign all remaining nonempty coalition values arbitrarily, with $v(\varnothing) = 0$. Every proposed observation is therefore produced by a game: O_{end} is surjective and has rank $2n+1$. There are $2^n - 1$ freely assignable nonempty coalition worths, so rank–nullity gives $\dim \ker O_{\mathrm{end}} = 2^n - 2n - 2$, and the first isomorphism theorem gives the stated quotient. □

The rank calculation counts what the scorer sees. The next criterion identifies which targets survive the changes that remain hidden.

A change of game is invisible to the scorer when it leaves every observation coordinate unchanged. A linear target is recoverable exactly when every such invisible change also leaves the target unchanged. The two kernels below express this test.

Lemma 10 (Identification criterion). *Let $\mathcal{G} \subseteq \Gamma^{\mathcal{N}}$ be a linear subspace, let W be a vector space, and let $L : \mathcal{G} \to W$ be linear. Then L factors linearly through $O_{\mathrm{end}}|_{\mathcal{G}}$ if and only if*

$$\ker(O_{\mathrm{end}}|_{\mathcal{G}}) \subseteq \ker(L|_{\mathcal{G}}).$$

Proof of Lemma 10. If $L = \ell \circ O_{\mathrm{end}}$, then an invisible game change v has $O_{\mathrm{end}}(v) = 0$ and hence $L(v) = 0$. Conversely, suppose the kernel containment holds. Define $\ell(O_{\mathrm{end}}(v)) = L(v)$ for each attainable observation. Two games with the same observation differ by an element of $\ker O_{\mathrm{end}}$, so they have the same L value. The definition is therefore unambiguous, and ℓ is linear because both maps are linear. □

A.2 ENDPOINT-LOCAL VALUES AND MIDPOINT SELECTION

Which requirements can an endpoint score retain when it cannot recover Shapley? Begin with the four axioms that characterize Shapley. *Efficiency* allocates the grand-coalition worth, so $\sum_i \phi_i(v) = v(\mathcal{N})$. *Anonymity* makes scores invariant to client renaming. *Linearity* makes the score of a sum of games equal the sum of their scores. The *null-player axiom* assigns zero to a client whose marginal worth is zero for every coalition. These four axioms characterize Shapley (Shapley, 1953).

For $n \geq 4$, the collision prevents any endpoint-local score from satisfying all four axioms on unrestricted games (Theorem 15). If we retain anonymity, linearity, and the null-player axiom while relaxing efficiency, total credit may differ from V. Adding *dummy calibration*, which assigns x to a client with constant marginal worth x, restricts every score to

$$\Psi_i^{(c)}(v) = (1-c)s_i + cm_i, \qquad c \in \mathbb{R}.$$

The proposition below derives this form. Even when the rule receives all clients' measurements, these axioms leave only the client's own two measurements. The choices $c = 0$ and $c = 1$ reward standalone and removal contributions, respectively; the segment $0 \leq c \leq 1$ belongs to classical semivalue theory (Dubey et al., 1981).

Giving the two perspectives equal status selects $c = 1/2$. Formally, the dual game $v^D(S) = V - v(\mathcal{N} \setminus S)$ swaps standalone and removal measurements. *Self-duality*, $\Psi(v^D) = \Psi(v)$, uniquely selects the midpoint within this family (Corollary 12). This explains its linear attribution properties; the smooth analysis and experiments assess its approximation error.

Proposition 11 (Linear endpoint-local values). *Let $n \geq 4$. Every linear, anonymous, null-player, endpoint-local value has constants $a, b \in \mathbb{R}$ such that*

$$\Psi_i(v) = bs_i + am_i.$$

If dummy calibration is imposed, $a + b = 1$, so for some $c \in \mathbb{R}$,

$$\Psi_i^{(c)}(v) = (1 - c)s_i + cm_i. \tag{10}$$

Proof of Proposition 11. We first show that an endpoint-local linear rule is a linear function of its measurements. Surjectivity of O_{end} defines ψ on every endpoint vector. For observations $x = O_{\text{end}}(v)$ and $y = O_{\text{end}}(w)$,

$$\psi(ax + by) = \Psi(av + bw) = a\psi(x) + b\psi(y),$$

so ψ is linear. Next fix player i. Renaming the other clients cannot change i's score. Their standalone measurements must therefore have one common coefficient, and their removal measurements another. Including V and i's own measurements leaves only five coefficients:

$$\Psi_i(v) = AV + Bs_i + C\sum_{j\neq i} s_j + Dm_i + E\sum_{j\neq i} m_j.$$

Finally make i null while leaving the other clients' game arbitrary: for a normalized game u on $\mathcal{N} \setminus \{i\}$, set $v(S) = u(S \setminus \{i\})$. Its score must be zero and $s_i = m_i = 0$. The remaining terms must vanish however the other clients interact. Because $n - 1 \geq 3$, Proposition 9 makes the remaining endpoint coordinates independently variable. Equivalently, the aggregate feature triple

$$\left(V, \sum_{j\neq i} s_j, \sum_{j\neq i} m_j\right)$$

can take three linearly independent values $(1, 0, n - 1)$, $(0, 1, 0)$, and $(0, 0, -1)$ under suitable basis games. The null-player requirement $\Psi_i(v) = 0$ for every such extension forces $A = C = E = 0$. Writing $b = B$ and $a = D$ leaves $bs_i + am_i$. If i has constant marginal contribution x, then $s_i = m_i = x$; dummy calibration for every x gives $a + b = 1$. □

The decisive step is varying the other clients' game while i stays null. Any surviving cross-client correction would give i nonzero credit in some such game. Shapley avoids this restriction by using coalition worths outside the endpoint observation; it can then satisfy all four classical axioms.

A.2.1 Monotonicity and self-duality

Two natural requirements narrow this family: increasing a measurement should never lower the score, and swapping the two perspectives should leave it unchanged. Start with the dummy-calibrated family,

$$\Psi_i^{(c)} = (1 - c)s_i + cm_i.$$

Holding all other observation coordinates fixed, monotonicity in s_i and m_i is equivalent to $1 - c \geq 0$ and $c \geq 0$, respectively. Surjectivity ensures these coordinate comparisons exist on the unrestricted TU space, so the condition is exactly $c \in [0, 1]$.

Corollary 12 (Self-dual selection). *Within the family in Equation* (10)*, self-duality $\Psi(v^D) = \Psi(v)$ uniquely selects the midpoint $\Psi_i^{1/2}(v) = (s_i + m_i)/2$ in Equation* (5)*.*

Proof of Corollary 12. For the dual game,

$$s_i(v^D) = v(\mathcal{N}) - v(\mathcal{N} \setminus \{i\}) = m_i(v), \qquad m_i(v^D) = s_i(v).$$

Consequently self-duality requires $(1 - 2c)(s_i - m_i) = 0$ for every game and player. The endpoint difference can be nonzero, so $c = 1/2$ is necessary; direct substitution proves sufficiency. □

A.3 Interaction error and pairwise recovery

To compare an observation-based score with Shapley, express both as sums of the same interaction effects. Shapley's ordering average is

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

The dividend decomposition in Section 4.1 is obtained by Möbius inversion:

$$d_v(T) = \sum_{S \subseteq T} (-1)^{|T|-|S|} v(S).$$

These formulas give the equivalent marginal and interaction representations.

Proposition 13 (Exact interaction error). *For every $c \in \mathbb{R}$,*

$$\Psi_i^{(c)}(v) - \phi_i(v) = \sum_{\substack{T \ni i \\ |T| \geq 2}} \left(c - \frac{1}{|T|}\right) d_v(T), \tag{11}$$

$$\sum_i \Psi_i^{(c)}(v) - v(\mathcal{N}) = \sum_{|T| \geq 2} (c|T| - 1) d_v(T). \tag{12}$$

Proof of Proposition 13. Möbius inversion gives $s_i = d_v(\{i\})$. Removing i from the grand coalition removes exactly the dividends indexed by coalitions containing i, so

$$m_i = \sum_{T \ni i} d_v(T).$$

Substitution into Equation (10) and coefficient-wise comparison with Shapley proves Equation (11). Summing over players counts an order-q dividend q times. Since Shapley is efficient, the aggregate difference has coefficient $cq - 1$, proving Equation (12). At $c = 1/2$, the coefficients vanish through order two and are strictly positive at every higher order. □

Corollary 14 (Pairwise compatibility diagnostic). *For $n \geq 3$, an endpoint observation has a signed 2-additive extension if and only if*

$$\sum_i \Psi_i^{1/2} = V.$$

This condition tests compatibility with the class; it does not imply that the unobserved game generating the endpoints is 2-additive.

Proof of the pairwise compatibility diagnostic. For any 2-additive extension, Proposition 2 and Shapley efficiency give $\sum_i \Psi_i^{1/2} = V$. Conversely, suppose this equality holds. Set the singleton dividends to s_i and let $r_i = m_i - s_i$ be the required incident pair mass. The equality is equivalent to $\sum_i r_i = 2(V - \sum_i s_i)$. Signed weights on the complete graph can realize any degree vector when $n \geq 3$: assign the desired degrees of clients $4, \ldots, n$ to their edges with client 1, reduce client 1's remaining degree accordingly, and realize the three remaining degrees on the triangle $\{1, 2, 3\}$ using edge weights $(d_1 + d_2 - d_3)/2$, $(d_1 + d_3 - d_2)/2$, and $(d_2 + d_3 - d_1)/2$. These pair dividends give every client incident mass r_i, and their total is $V - \sum_i s_i$. The resulting 2-additive game therefore has the prescribed standalone gains, removal losses, and grand-coalition worth. Higher-order signed dividends can preserve the same equality, so it certifies only the existence of a pairwise extension. □

A.4 Order threshold and fixed-evaluator witness

Proof of Theorem 3. We move effect between two pairs and two triples so that each client's total stays fixed. Start with dividend one on every nonempty coalition of order at most three and zero otherwise. For $0 < \epsilon < 1$, define the perturbation η by

$$\eta(\{1,3\}) = \epsilon, \qquad \eta(\{1,4\}) = -\epsilon, \qquad \eta(\{1,2,3\}) = -\epsilon, \qquad \eta(\{1,2,4\}) = \epsilon,$$

with $\eta(T) = 0$ for every other coalition. Let v be the baseline game and let w have dividends $d_w(T) = d_v(T) + \eta(T)$. Their total perturbation is zero. Their incidence sums for players one through four are, respectively,

$$\epsilon - \epsilon - \epsilon + \epsilon = 0, \quad -\epsilon + \epsilon = 0, \quad \epsilon - \epsilon = 0, \quad -\epsilon + \epsilon = 0.$$

All other incidence sums are zero. Thus the grand-coalition worth, standalone measurements, and leave-one-out measurements are unchanged, so $O_{\mathrm{end}}(v) = O_{\mathrm{end}}(w)$. All perturbed dividends remain nonnegative, and both games are 3-additive. Dividing pair perturbations by two and triple perturbations by three gives Shapley displacement $(0, 0, \epsilon/6, -\epsilon/6, 0, \ldots, 0)$. □

Theorem 15 (Classical axioms require additional game information). *For $n \geq 4$, no endpoint-local value on all normalized TU games satisfies efficiency, anonymity, the null-player axiom, and linearity simultaneously.*

Proof of Theorem 15. Efficiency, anonymity, the null-player axiom, and linearity characterize Shapley. Theorem 3 gives two complete games compatible with one endpoint observation and different Shapley values, so Shapley lies outside the endpoint-local family. □

Proof of Corollary 4. For every 2-additive game, Equation (11) at $c = 1/2$ gives $\phi = \Psi^{1/2}$, so the endpoint observation recovers Shapley. For every $k \geq 3$, the witness difference d lies in the k-additive subspace, satisfies $d \in \ker O_{\mathrm{end}}$, and has $\phi(d) \neq 0$. Lemma 10 completes the threshold proof. □

A.4.1 Fixed-evaluator realization

Proof of Proposition 5. We need one utility function to reproduce both games on different update profiles while giving those profiles the same full aggregate. Let v, w be the positive games in Theorem 3, and replace each dividend by the corresponding multilinear monomial to obtain f_v, f_w. Their difference is

$$f_w(x) - f_v(x) = \epsilon x_1(1 - x_2)(x_3 - x_4),$$

so $f_v(t\mathbf{1}) = f_w(t\mathbf{1})$ for every $t \in \mathbb{R}$.

Write points of $\mathbb{R}^{2n}$ as (x, y), let e_i be coordinate i of the standard basis of $\mathbb{R}^n$, and define the weighted update components

$$z_i = (e_i, 0), \qquad z_i' = \left(\frac{1}{n}\mathbf{1}, e_i - \frac{1}{n}\mathbf{1}\right).$$

They have the same full sum, $\sum_i z_i = \sum_i z_i' = (\mathbf{1}, 0)$. For $\bar{x} = n^{-1}\sum_j x_j$, set $\theta = 0$ and define the shared utility

$$F(x, y) = f_v(x) + f_w(\bar{x}\mathbf{1} + y) - f_v(\bar{x}\mathbf{1}).$$

This polynomial has degree at most three. Its correction terms switch which game is evaluated: for the first profile they cancel, whereas for the second they replace f_v by f_w. In detail, at $(\mathbf{1}_S, 0)$, diagonal agreement gives

$$F(\mathbf{1}_S, 0) = f_v(\mathbf{1}_S) = v(S).$$

At the second profile's coalition point, with $t = |S|/n$,

$$F(t\mathbf{1}, \mathbf{1}_S - t\mathbf{1}) = f_w(\mathbf{1}_S) = w(S).$$

Finally take $U_i = z_i/q_i$ and $U_i' = z_i'/q_i$. Both games therefore share the checkpoint, utility function, positive weights, released aggregate, and endpoint observation. Their Shapley values differ. □

Corollary 16 (Quantitative endpoint ambiguity). *Let $\widehat{\phi}$ be any deterministic estimator based on the augmented observation $x = (O_{\mathrm{end}}, \theta, F, (q_i), \sum_i q_i U_i)$. For the two profiles in Proposition 5 and any norm,*

$$\max\{\|\widehat{\phi}(x) - \phi(v)\|, \|\widehat{\phi}(x) - \phi(w)\|\} \geq \tfrac{1}{2}\|\phi(v) - \phi(w)\|.$$

For the witness parameter ϵ, this lower bound is $\epsilon/12$ in ℓ_∞ and $\epsilon/6$ in ℓ_1.

Proof of Corollary 16. The common augmented observation forces the same estimate $a = \widehat{\phi}(x)$ for both profiles. The triangle inequality gives

$$\|\phi(v) - \phi(w)\| \le \|\phi(v) - a\| + \|a - \phi(w)\|,$$

so at least one error is at least half the left-hand side. Substituting the displacement $(0, 0, \epsilon/6, -\epsilon/6, 0, \dots, 0)$ gives the stated norms. □

Public individual updates can resolve the ambiguity in this lift. A shared quadratic utility under additive fixed directions induces interactions of order at most two and therefore excludes the order-three witness.

A.4.2 EDGE CASES

With at most three clients, the endpoint observation reveals every nonempty coalition worth. For one or two clients some coordinates repeat; for three, the singletons, leave-one-out coalitions, and grand coalition are the seven nonempty coalitions. There is therefore no hidden coalition worth to vary. Four clients are also the minimum for the representation proof: after making one client null, three must remain so that their endpoint coordinates can vary independently.

A.5 SHARP POSITIVE IDENTIFIED SETS

The example leaves a range of possible scores for each client. We now compute that range over all positive 3-additive games with the same observation. Their Shapley vectors form the *identified set*.

First remove the known standalone effects. The remaining total worth is $J = V - \sum_i s_i$, and client i's remaining incident effect is $r_i = m_i - s_i$. Let P and H denote the total pair and triple effects. Their sum is J. Summing the r_i counts every pair twice and every triple three times, so

$$D = \sum_i r_i = 2P + 3H, \qquad P = 3J - D, \qquad H = D - 2J.$$

The observation therefore determines the total effect at each order. What remains unknown is how much triple effect involves each client. Denote this incident triple mass by t_i. Shapley splits pair effects in half and triple effects in thirds, giving

$$\phi_i = s_i + \frac{r_i - t_i}{2} + \frac{t_i}{3} = \Psi_i^{1/2} - \frac{t_i}{6}.$$

Thus bounding t_i bounds the score. Positivity requires $0 \le t_i \le H$ and $0 \le r_i - t_i \le P$; counting triple effects over clients also gives $\sum_i t_i = 3H$. The theorem makes these constraints explicit and shows that they suffice to construct compatible games.

The proof has two tasks: realize every allowed pair/triple assignment as a game, then find the largest and smallest possible t_i. Realizability is what makes the bounds sharp rather than merely necessary.

Proof of Theorem 6. Let P and H denote the total pair- and triple-dividend masses of a positive 3-additive extension. The endpoint equations imply

$$J = P + H, \qquad D = 2P + 3H,$$

and hence $P = 3J - D$ and $H = D - 2J$. For each player, write

$$t_i = \sum_{T \ni i,\ |T|=3} d_v(T), \qquad p_i = \sum_{T \ni i,\ |T|=2} d_v(T) = r_i - t_i.$$

Pair dividends can be viewed as nonnegative edge weights: p_i is the total weight incident to client i. With total edge weight P, the possible incident-sum vectors are obtained by distributing that total over pairs. Putting all weight on pair T gives $P\mathbf{1}_T$; splitting the weight gives convex combinations of these vectors. Conversely, an incident-sum vector within the coordinate bounds and with the correct total is always such a combination. After normalizing a positive total, a nonintegral candidate vertex

has two fractional coordinates that can move in opposite directions, so only pair indicators can be vertices. A zero total gives the zero vector. Thus

$$P\,\mathrm{conv}\{\mathbf{1}_T : |T| = 2\} = \left\{ p : 0 \le p_i \le P,\ \sum_i p_i = 2P \right\}.$$

The same argument for nonnegative triple weights with total mass H gives

$$H\,\mathrm{conv}\{\mathbf{1}_T : |T| = 3\} = \left\{ t : 0 \le t_i \le H,\ \sum_i t_i = 3H \right\}.$$

Substituting $p_i = r_i - t_i$ reduces extension feasibility to

$$t \in \prod_i [\ell_i, u_i], \qquad \sum_i t_i = 3H.$$

A vector can satisfy these interval and total constraints exactly when all intervals are nonempty and their lower and upper sums bracket $3H$. Every such t can be realized by triple dividends, and $p = r - t$ can be realized independently by pair dividends. Mapping t to $\Psi^{1/2} - t/6$ therefore gives every compatible Shapley vector. To bound coordinate i, let the other coordinates absorb as much or as little of $3H$ as their intervals permit. They attain every intermediate sum, giving precisely $[t_i^-, t_i^+]$. Finally,

$$\phi_i = s_i + \frac{p_i}{2} + \frac{t_i}{3} = s_i + \frac{r_i}{2} - \frac{t_i}{6}.$$

The map is decreasing in t_i, so the projected endpoints give the stated Shapley bounds. The convex-combination representations supply nonnegative pair and triple dividends attaining each coordinate endpoint, proving sharpness. □

The characterization covers degenerate cases. If $H = 0$, every $t_i = 0$ and the class reduces to positive 2-additive extensions. If $P = 0$, then $t_i = r_i$ and the remaining constraints are $0 \le r_i \le H$ and $\sum_i r_i = 3H$. For $n = 3$, the endpoint observation determines the complete game and the coordinate intervals necessarily collapse. Non-identification first becomes possible at $n = 4$.

A.5.1 Joint feasibility and sharp rankings

The coordinate intervals alone do not allow arbitrary combinations of scores: their total must be V. We first show that this is the only extra constraint, then use it to compare two clients. Let $[a_i, b_i]$ denote the sharp coordinate intervals. Every compatible Shapley vector lies in their Cartesian product and sums to V. Conversely, any vector x satisfying these two conditions gives $t = 6(\Psi^{1/2} - x)$ within the tightened triple intervals, hence within $\prod_i [\ell_i, u_i]$. Since $\sum_i \Psi_i^{1/2} - V = H/2$, it also satisfies $\sum_i t_i = 3H$. Equation (7) therefore gives

$$\mathcal{I}(o) = \{x : a_i \le x_i \le b_i,\ \sum_i x_i = V\}.$$

For distinct clients i, j, the lower endpoint of $\phi_i - \phi_j$ is $a_i - b_j$, and the upper endpoint is $b_i - a_j$. To prove attainability of the lower endpoint, take a feasible x with $x_i = a_i$. The remaining coordinates can realize $x_j = b_j$ exactly when

$$\sum_{k \ne i,j} a_k \le V - a_i - b_j \le \sum_{k \ne i,j} b_k.$$

The left inequality follows from feasibility of the coordinate endpoint $x_j = b_j$; the right follows from feasibility of $x_i = a_i$. The remaining intervals attain every intervening sum, so the two opposing endpoints are jointly attainable. Interchanging i, j proves the other endpoint; convexity gives the full interval. Thus $a_i > b_j$ is necessary and sufficient for a strict ordering in all compatible games. When $a_i = b_j$, a compatible tie exists. This ranking statement uses the same positive third-order assumption as the identified set.

A.5.2 Unbounded compatible scores without sign restrictions

Theorem 3 supplies the direction needed for the unboundedness claim in Section 4.3. Let $d = w - v$ be its game difference. Then $O_{\mathrm{end}}(d) = 0$ while $\phi(d) \neq 0$. For any game g and any real λ, the game $g + \lambda d$ has the same endpoint observation as g, whereas

$$\phi(g + \lambda d) = \phi(g) + \lambda\phi(d).$$

At least one coordinate is therefore unbounded as λ varies. This argument permits signed effects and imposes no fixed utility range.

B Smooth approximation and certification

This section proves the client-specific curvature certificate and the matching uniform error order. It also records how the relative rate changes when the leading Shapley signal vanishes.

Fix a checkpoint θ, a utility $F : \mathbb{R}^d \to \mathbb{R}$, and weighted update directions $z_i = q_i U_i$. Write $z_S = \sum_{i \in S} z_i$ and scale all updates by the same ε to obtain $v_\varepsilon(S) = F(\theta + \varepsilon z_S) - F(\theta)$. The parameter ε changes only the common update scale. We use a fixed norm on $\mathbb{R}^d$ and its induced multilinear operator norms.

Proof of Theorem 7. We first justify the integral and curvature statements used in the main text. For any game, expanding the expectation in Equation (8) gives

$$\int_0^1 f_i(t)\,dt = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \big(v(S \cup \{i\}) - v(S)\big) \int_0^1 t^{|S|}(1-t)^{n-1-|S|}\,dt.$$

The beta integral equals $|S|!(n-1-|S|)!/n!$, which is the Shapley coefficient. Twice integrating by parts also gives

$$\frac{f_i(0) + f_i(1)}{2} - \int_0^1 f_i(t)\,dt = \frac{1}{2}\int_0^1 t(1-t) f_i''(t)\,dt.$$

This proves Equation (9) because $\int_0^1 t(1-t)\,dt = 1/6$.

We now bound the curvature using information that a protocol can supply. Let Δ_j denote a discrete marginal difference. Differentiating the multilinear expectation twice yields

$$f_i''(t) = 2 \sum_{\substack{j<k \\ j,k \neq i}} \mathbb{E}\Big[\Delta_i \Delta_j \Delta_k v_\varepsilon(S_t^{ijk})\Big],$$

where S_t^{ijk} independently includes each client outside $\{i, j, k\}$ with probability t. Each third difference is an integral over the three update directions:

$$\Delta_i \Delta_j \Delta_k v_\varepsilon(S) = \int_{[0,\varepsilon]^3} D^3 F(\theta + \varepsilon z_S + t_i z_i + t_j z_j + t_k z_k)[z_i, z_j, z_k]\,dt_i\,dt_j\,dt_k.$$

The integration points lie in K by convexity. Thus its absolute value is at most $L\varepsilon^3 a_i a_j a_k$, and

$$\sup_{t \in [0,1]} |f_i''(t)| \leq 2L\varepsilon^3 a_i \sum_{\substack{j<k \\ j,k \neq i}} a_j a_k.$$

Substitution into the trapezoidal remainder proves the client-specific bound.

To identify the leading error rather than only bound its size, keep the cubic Taylor term. Define

$$Q_\varepsilon(S) = \varepsilon DF(\theta)[z_S] + \frac{\varepsilon^2}{2} D^2 F(\theta)[z_S, z_S].$$

This game contains only singleton and pair dividends, so the midpoint and Shapley agree on it. Continuity of $D^3 F$ gives the expansion for every coalition:

$$v_\varepsilon(S) = Q_\varepsilon(S) + \frac{\varepsilon^3}{6} D^3 F(\theta)[z_S, z_S, z_S] + o(\varepsilon^3).$$

Define the cubic game $C(S) = D^3F(\theta)[z_S, z_S, z_S]/6$. Its dividend on a distinct triple $\{i, j, k\}$ is $D^3F(\theta)[z_i, z_j, z_k]$, and every dividend above order three equals zero. The actual Taylor term is $\varepsilon^3 C$. Equation (11) leaves coefficient $1/6$ on precisely the triples containing i, proving the expansion. Finally, $\phi_i(v_\varepsilon) = \varepsilon DF(\theta)[z_i] + O(\varepsilon^2)$. When this first-order vector is nonzero, its ℓ_1 norm is $\Theta(\varepsilon)$. The numerator is $O(\varepsilon^3)$. □

The C^3 assumption supports the cubic remainder. An order-q dividend hierarchy of $O(\varepsilon^q)$ would require corresponding higher derivatives. Second-order Taylor approximations also appear in parameter-player Shapley analysis (Xu et al., 2025); here, exactness on the quadratic client-update game yields the explicit midpoint–Shapley error term.

B.1 Proof of the matching local lower bound

Proof of Theorem 8. The earlier collision must now persist as updates shrink, without changing the utility function. We achieve this by making every monomial cubic: common scaling then multiplies the whole collision by ε^3. Fix $n \geq 4$, write $\rho = 1/2$, and define

$$f_v(x) = \sum_i x_i^3 + \sum_{i<j} x_i^2 x_j + \sum_{i<j<k} x_i x_j x_k, \qquad f_w(x) = f_v(x) + \rho x_1(x_3 - x_4)(x_1 - x_2).$$

These polynomials still reproduce the original games v, w at coalition indicator vectors. On those zero-one inputs, $x_1^2 = x_1$, giving the same four dividend perturbations. Their dividends through order three are nonnegative, their endpoint observations agree, and $\phi(w) - \phi(v) = (e_3 - e_4)/12$. Moreover, $f_v(t\mathbf{1}) = f_w(t\mathbf{1})$ for all real t.

Use the fixed weighted directions from Proposition 5,

$$z_i = (e_i, 0), \qquad z_i' = (\mathbf{1}/n, e_i - \mathbf{1}/n), \qquad U_i = z_i/q_i, \quad U_i' = z_i'/q_i.$$

Add a shared linear term so that the Shapley vector has a nonzero first-order signal while its hidden difference remains cubic. With $\theta = 0 \in \mathbb{R}^{2n}$ and $\bar{x} = n^{-1}\sum_i x_i$, set

$$F(x, y) = \sum_i x_i + f_v(x) + f_w(\bar{x}\mathbf{1} + y) - f_v(\bar{x}\mathbf{1}).$$

Diagonal agreement and homogeneity imply, for every coalition S,

$$v_\varepsilon(S) = F(\varepsilon \mathbf{1}_S, 0) = \varepsilon|S| + \varepsilon^3 v(S), \qquad w_\varepsilon(S) = \varepsilon|S| + \varepsilon^3 w(S).$$

The second identity evaluates F at $(\varepsilon|S|\mathbf{1}/n, \varepsilon(\mathbf{1}_S - |S|\mathbf{1}/n))$. Both games are positive and 3-additive. Linearity of the observation gives $O_{\mathrm{end}}(v_\varepsilon) = O_{\mathrm{end}}(w_\varepsilon)$ for every scale, and both released aggregates equal $\varepsilon(\mathbf{1}, 0)$. The checkpoint, evaluator, weights, and hidden profiles are fixed throughout this family.

Shapley linearity gives

$$\phi(w_\varepsilon) - \phi(v_\varepsilon) = \frac{\varepsilon^3}{12}(e_3 - e_4).$$

The estimator sees the same input under both profiles, even when told ε. Its output therefore has one common distribution. Use a single draw Y from that distribution to compare both possible targets. For every draw, the two errors together cover at least the targets' separation; taking expectations gives

$$\mathbb{E}\|Y - \phi(v_\varepsilon)\|_1 + \mathbb{E}\|Y - \phi(w_\varepsilon)\|_1 \geq \|\phi(v_\varepsilon) - \phi(w_\varepsilon)\|_1 = \varepsilon^3/6.$$

At least one expected error is $\varepsilon^3/12$ or larger. The argument includes deterministic rules; it bounds expectation rather than the probability of an error. To obtain relative error, note that both profiles have first-order Shapley vector $\mathbf{1}$. Positivity and efficiency give the same target norm:

$$\|\phi(v_\varepsilon)\|_1 = \|\phi(w_\varepsilon)\|_1 = n\varepsilon + C_n\varepsilon^3, \qquad C_n = n + \binom{n}{2} + \binom{n}{3}.$$

The relative lower bound is therefore $\varepsilon^2/[12(n + C_n\varepsilon^2)]$, which is at least $c_n\varepsilon^2$ for $c_n = [12(n + C_n)]^{-1}$ and $0 < \varepsilon \leq 1$. Finally, the convex hull of the two finite families of coalition points at unit scale and the origin contains every segment at smaller scales. It is compact, and the fixed polynomial F has bounded derivatives there. Both profiles thus satisfy the smooth theorem on one fixed region with finite constants independent of ε. □

The construction is a worst-case obstruction within smooth fixed-update games. It does not assert that a prescribed trained architecture realizes these profiles, nor that its ambiguity attains the lower bound.

B.2 Relative rates when the leading signal vanishes

Cubic absolute error can be small compared with one Shapley vector and large compared with another. The relative rate depends on the first nonzero term of that vector. The expansion below makes the condition explicit.

Corollary 17 (Signal-dependent relative rates). *Under the assumptions of Theorem 7, define*

$$a_i = DF(\theta)[z_i], \qquad b_i = \tfrac{1}{2}D^2F(\theta)[z_i, z_{\mathcal{N}}], \qquad c = \phi(C), \quad C(S) = \tfrac{1}{6}D^3F(\theta)[z_S, z_S, z_S].$$

For fixed profiles, $\phi(v_\varepsilon) = \varepsilon a + \varepsilon^2 b + \varepsilon^3 c + o(\varepsilon^3)$, and the midpoint's unregularized relative ℓ_1 error has the following upper orders:

Condition	$\|\phi(v_\varepsilon)\|_1$	$\|\Psi^{1/2}(v_\varepsilon) - \phi(v_\varepsilon)\|_1 / \|\phi(v_\varepsilon)\|_1$
$a \neq 0$	$\Theta(\varepsilon)$	$O(\varepsilon^2)$
$a = 0,\ b \neq 0$	$\Theta(\varepsilon^2)$	$O(\varepsilon)$
$a = b = 0,\ c \neq 0$	$\Theta(\varepsilon^3)$	$O(1)$

Proof. Apply Shapley to each term of the Taylor expansion from Appendix B. Linearity allows this term-by-term calculation. The linear game gives a. In the quadratic game $D^2F(\theta)[z_S, z_S]/2$, client i receives its singleton term $D^2F(\theta)[z_i, z_i]/2$ and half each incident pair term $D^2F(\theta)[z_i, z_j]$, giving b_i. The cubic game gives c. The first nonzero coefficient determines the norm order; dividing the $O(\varepsilon^3)$ absolute bound proves all three upper rates.

□

A stationary checkpoint has $DF(\theta) = 0$, hence $a = 0$; its quadratic coefficient must then be checked before interpreting relative error. The ratios use no denominator floor.

B.3 Role of client weights

Substituting $z_i = q_i U_i$ into Theorem 7 gives

$$\Psi_i^{1/2}(v_\varepsilon) - \phi_i(v_\varepsilon) = \frac{\varepsilon^3}{6} \sum_{\substack{j<k \\ j,k \neq i}} q_i q_j q_k D^3F(\theta)[U_i, U_j, U_k] + o(\varepsilon^3).$$

This expression shows where client weights and update geometry enter the error. Neither data heterogeneity nor update dispersion alone determines it: the utility's third derivative and the weighted update directions act together. The public caps used by the certificate bound this contraction without exposing those directions.

C Verification of the scaling law and certificate

This section gives the protocol needed to reconstruct the two empirical claims: the common-scale error law and the decision certificates in controlled logistic rounds. Complete coalition enumeration is used only to obtain exact Shapley values for evaluation; the midpoint and certificates receive the information stated in the main text. For the scaling experiment, we also compute the fraction of absolute dividend mass due to groups of at least three clients:

$$H_{\geq 3} = \frac{\sum_{|T| \geq 3} |d_v(T)|}{\sum_{T \neq \varnothing} |d_v(T)| + 10^{-12}}.$$

This quantity describes the evaluated game; it is not an ambiguity measure over other games with the same endpoint observation.

Table 2: Reproducibility settings for the experiments used in the main text. All games use fixed, sample-size client weights. GD denotes deterministic full-batch gradient descent.

Collection	Games	Model and utility	Update construction
UCI-HAR development	20	Logistic regression; held-out negative cross-entropy	Checkpoint GD: 100 steps, lr 1, $\ell_2 = 10^{-4}$; local GD: 5 steps, lr .5, $\ell_2 = 10^{-4}$; clip .2; $\varepsilon = .5$
MHEALTH confirmation	20	Same model and utility on a held-out volunteer	Same optimizer, clip, scale, certificate, and decision gates as UCI-HAR
Fashion-MNIST scaling	30 profiles	49–24–10 tanh; negative Brier utility	Local GD: 20 steps, lr .1; each fixed profile evaluated at five common scales

C.1 Bounded-logistic curvature certificate

For multinomial logistic parameters W and an augmented feature x, one example's negative cross-entropy utility is

$$h_y(Wx) = (Wx)_y - \log \sum_c \exp((Wx)_c).$$

Its third derivative is a categorical third central moment. The categorical covariance and direction-range bounds give

$$|D^3 h_y(r)[a,b,c]| \le \frac{1}{\sqrt{2}} \|a\|_2 \|b\|_2 \|c\|_2.$$

Unit-norm features therefore give $L = 1/\sqrt{2}$ in Frobenius norm. If local updates are clipped at b before applying public weight q_i, then $a_i = q_i b$ and the scorer constructs

$$R_i = \frac{\varepsilon^3}{6\sqrt{2}}\, a_i \sum_{\substack{j<k \\ j,k\neq i}} a_j a_k, \qquad [\ell_i, u_i] = [\Psi_i^{1/2} - R_i, \Psi_i^{1/2} + R_i].$$

Intersecting this box with efficiency tightens each coordinate to

$$\ell_i' = \max\left\{ \ell_i, V - \sum_{j\neq i} u_j \right\}, \qquad u_i' = \min\left\{ u_i, V - \sum_{j\neq i} \ell_j \right\}.$$

For UCI-HAR (Reyes-Ortiz et al., 2013), each of 20 deterministic assignments uses 13 official-training subjects for the checkpoint, eight as clients, and the official test subjects for evaluation. Features are standardized using checkpoint data, augmented with a bias coordinate, and projected to norm at most one. For MHEALTH (Banos et al., 2014), we exclude the null activity, retain every tenth labeled row, and apply the same checkpoint-only standardization, bias augmentation, and projection. Each of 20 assignments uses one volunteer for the checkpoint, eight as clients, and one for evaluation.

We selected $\varepsilon = .5$ on UCI-HAR, then fixed the optimizer settings in Table 2, certificate, and decision gates before evaluating MHEALTH. The scorer receives the feature bound, clip, scale, public weights, and endpoint observation, but no update directions, evaluation outcomes, additional coalition values, or Shapley values. Full enumeration of the 2^8 coalitions is used only to audit coverage and decisions.

Across the 20 MHEALTH games, the intervals cover every exact Shapley vector, have median relative width 1.24%, certify 159/160 signs and 555/560 strict pairs, and identify all 20 top clients. All 160 updates reach the clip, so the experiment tests an actively controlled protocol. Because the assignments reuse ten volunteers, the result establishes transfer to a separate dataset, not prevalence across independent populations.

Figure 3 makes the decision criteria concrete in the first indexed confirmation game. The example is selected by its fixed profile index rather than by its outcome.

C.2 Common-scaling experiment

For Fashion-MNIST, we average-pool the canonical images to 7×7, standardize using 12,000 training examples, and evaluate utility on a fixed stratified set of 500 examples. The model has

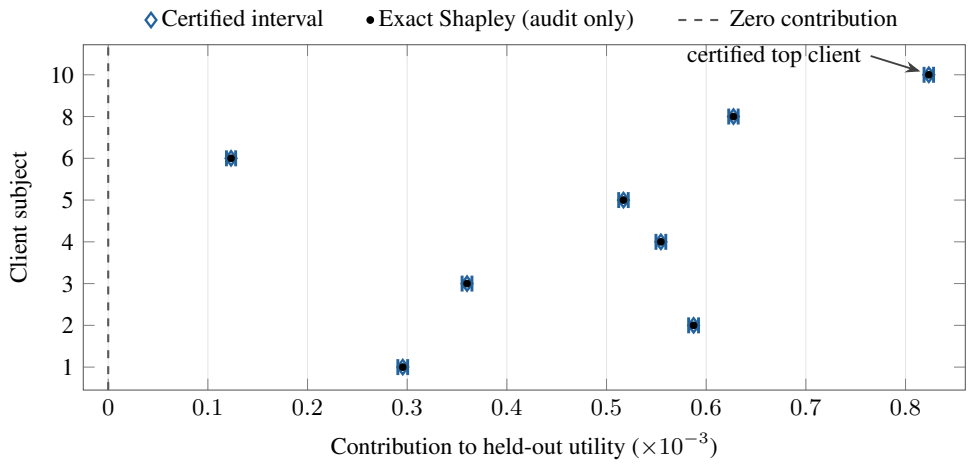


Figure 3: Client-level certificates in the first indexed MHEALTH confirmation game at $\varepsilon = .5$. Horizontal bars are the efficiency-constrained intervals constructed from the endpoint observation and public controls; open diamonds mark their endpoint midpoints. Black dots are exact Shapley values from complete coalition enumeration and are used only to audit the certificate. All eight intervals exclude zero and are pairwise disjoint, while client subject 10's lower endpoint exceeds every other upper endpoint. This one game therefore certifies every sign, all 28 strict pairwise rankings, and its top client.

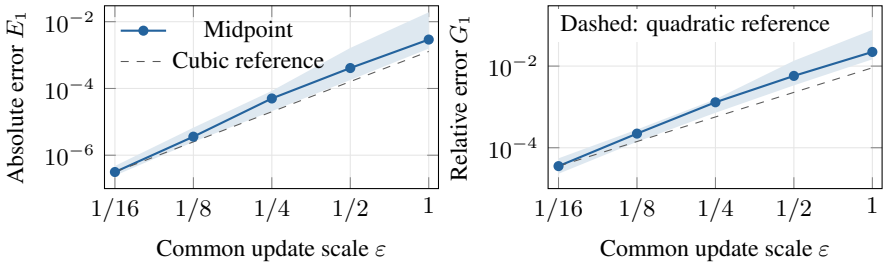


Figure 4: Median midpoint error for 30 fixed Fashion-MNIST profiles; shading is the interquartile range. Dashed order-three and order-two references are anchored at $1/16$, not fitted bounds. Fits at $\varepsilon \leq 1/2$ have median slopes 3.125 and 2.110. At each scale, the midpoint uses only each client's two measurements.

one 24-unit tanh hidden layer followed by softmax; its parameter-to-Brier utility is smooth. Five partition seeds, two checkpoints, and three Dirichlet concentrations give 30 profiles. Each profile uses twenty local full-batch steps and is held fixed while we multiply every update by five common scales. The fitted slopes use the four scales at most $1/2$.

The multiclass Brier loss is

$$L_{\text{Brier}}(p, y) = \sum_{c=1}^{10} (p_c - \mathbf{1}\{y = c\})^2 \in [0, 2].$$

We enumerate every coalition at every scale, eliminating coalition-sampling error; uncertainty from the finite evaluation set remains. The unscaled sum $\sum_i \|q_i U_i\|_2$ ranges from .865 to 1.648; every fitted gap exceeds the 10^{-14} numerical floor. Across the fitted scales, the smallest $\|\phi\|_1$ is .005179, over 5×10^9 times the 10^{-12} denominator offset used in the empirical relative-error metric. Profile-level absolute slopes range from 2.912 to 4.643: faster decay is allowed when the leading cubic contraction vanishes.

C.3 Scope of the positive third-order model

Theorem 6 gives exact intervals only when the true game has nonnegative effects through order three. As a scope check, we applied its endpoint feasibility conditions to 273 retained fixed-weight negative-Brier games. Six observations (2.2%) admit such an extension, and the exact Shapley

Table 3: Closest information models and the distinction used in this paper.

Line of work	Observed information	Target	Difference here
Private FL scores (Pejó et al., 2021; Jaramillo Velez et al., 2026)	Standalone and removal evaluations	A deployable contribution score	Hold the same endpoints fixed and characterize Shapley identification and error
Partial-game values (Housman, 2002; Bok & Černý, 2024)	A declared partial characteristic function	Select an allocation or extension value	Require one rule to agree across every complete game sharing the observation
Positive and player-centered extensions (Černý & Grabisch, 2024; Černý et al., 2024; Černý, 2025)	General or focal-player known-coalition systems	Feasible extensions and solution ranges	Use all clients' endpoint coalitions and obtain a closed form under order-three positivity
Semivalues and low-order surrogates (Fried, 2026; Pelegrina et al., 2026)	Order-separated marginal values or chosen coalition samples	Recover information or fit an approximation	Use only the fixed endpoint observation, which merges all shared interaction orders

vector of each observed game falls outside at least one resulting interval. Compatibility of an observation with the class therefore does not establish that the observed game belongs to it.

This finding limits the empirical use of the structural intervals but not their identified-set characterization. It motivates the curvature certificate, which permits signed effects of arbitrary order and is the paper's primary empirical result.

D Relation to incomplete-game models

The closest lines of work differ in either what they observe or what they ask that information to determine. Table 3 makes those differences explicit.

A partial-game method selects an allocation or completion for one incomplete object (Housman, 2002; Bok & Černý, 2024; Černý & Grabisch, 2024). Our endpoint-local rule instead acts on complete games and must return the same value whenever their observations agree. Requiring the null-player axiom across all such games is what removes dependence on the grand-coalition worth and other clients' measurements in Proposition 11.

General positive-extension theory describes feasibility through nonnegative dividends (Černý et al., 2024; Černý, 2025). The endpoint and third-order restrictions let us eliminate the individual dividends and obtain closed-form intervals and rankings.

Observing every anonymous semivalue reveals each client's effect at every interaction order (Fried, 2026); observing only the standalone and removal measurements does not. Low-order surrogate fitting and Taylor approximations instead introduce additional modeling or coalition information (Pelegrina et al., 2026; Xu et al., 2025). Our results hold the endpoint information fixed and characterize what it determines exactly or approximately.