WHAT DETERMINES A LOCAL UPDATE MAP?

Anonymous authors
Paper under double-blind review

ABSTRACT

What information determines the result of local optimization? Two groupings of the same strictly convex quadratic losses can preserve the objective and complete separate gradient and Hessian marginals, yet produce different two-step updates. Their within-child pairing resolves this ambiguity. For quadratic children, agreement at the first q horizons identifies all later endpoints at that input, where q counts the actual union spectrum. Agreement at other inputs additionally depends on matrix-valued spectral weights: even an entire shared trajectory and matching eigenvalue distributions can leave the maps different elsewhere. We realize sharp information limits with partitions of the same atomic losses, including noncommuting children, and quantify the observation depth needed with noisy logs. Smooth remainder bounds and controlled logistic and trained-neural evaluations locate the range of leading-pairing prediction beyond quadratics. A ridge comparison illustrates a further distinction: covariance-aware recovery can resolve later-update comparisons that spectrum-only bounds leave undecided with the same early logs. The results separate information that controls the size of an update change from information that identifies the update itself.

1 INTRODUCTION

Local optimization averages parameters after several gradient steps within each group of losses (McMahan et al., 2017; Stich, 2019). Existing surrogate-map analyses characterize this map for quadratic losses (Charles & Konečný, 2021; Charles & Rush, 2022). We ask an identification question: which information shared by two groupings of the same losses forces their updates to coincide? We distinguish objective agreement, matching a reference trajectory, and identifying the map from derivative summaries or early logs, rather than complete child objectives and assignments.

We establish two boundaries. The first concerns derivative information. Two partitions of the same strictly convex quadratic losses can preserve the complete child-gradient multiset and the complete child-Hessian multiset at the input, while their two-step gaps from the unsplit update are zero and $\eta^2\sqrt{V_g V_H}$. The missing information is the within-child pairing $A_{\mathcal{P}} = \sum_k p_k H_k(g_k - \bar{g})$.

The second boundary concerns observations of the update itself. For quadratic children with q distinct Hessian eigenvalues, agreement at the first q horizons identifies all later endpoints at that input. No simultaneous diagonalization is needed. But agreement at one input need not identify the whole map: two partitions of the same two-dimensional losses can match every horizon along a nonzero reference trajectory, preserve the eigenvalue distribution, and still disagree at another input. Matrix-valued spectral weights supply the missing input dependence. With these weights fixed, we give a dimension-free sharp first-disagreement bound, attained by noncommuting children formed from the same atoms.

With noisy logs, we characterize the observation depth that uniformly limits the distance between compatible endpoints. The key constructive step is to realize the extremal information as two partitions of one strictly convex loss pool, including finite equal-mass examples with controlled atomic curvature. Smooth remainder bounds and logistic and trained-neural evaluations locate the short-horizon range of pairing prediction; frozen curvature separates composition from nonlinear error. A ridge example distinguishes spectrum-only and covariance-aware readouts of the same early logs: additional matrix information can resolve comparisons that the uniform bound leaves open.

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.

These questions arise when one participant's data are split across workers or storage shards. We fix the data, weights, input, work, and update rule, and vary their assignment to synchronized children. Full-batch gradient descent isolates this effect from sampling and persistent optimizer state.

Our contributions are:

1. A sharp same-data ambiguity under complete unpaired derivative marginals, identifying which pairing information they omit.
2. A finite-horizon identification boundary for noncommuting quadratic children, a sharp dimension-free ambiguity bound, and a same-data counterexample separating single-input from all-input testing.
3. A sharp noisy-observation diameter and its implied minimum uniform observation depth, with finite equal-mass realizations and explicit atomic-curvature bounds. Smooth remainder bounds and controlled learning examples distinguish information limits from achievable prediction.

Relation to closest work. The pairing is a signed local-update drift term. For quadratic children, with $x_k = w - \eta g_k$ and $\bar{x} = w - \eta\bar{g}$, the averaged-gradient discrepancy $\sum_k p_k \nabla F_k(x_k) - \nabla \bar{F}(\bar{x})$ is exactly $-\eta A_{\mathcal{P}}(w)$. Wang et al.'s pseudo-Lipschitz condition bounds this discrepancy (Wang et al., 2024, Eq. (9)); second-order heterogeneity controls related convergence terms (Patel et al., 2025; 2026). For the two constants evaluated in Appendix I.3, our matched constructions preserve their values while changing the realized update, separating worst-case control from identification of a particular map.

Our lower bounds require a simultaneous realization: the extremal moment data must come from two partitions of one strictly convex loss pool while preserving the specified derivative or spectral information. We construct these matched alternatives with noncommuting children and finite equal-mass realizations with controlled atomic curvature. The proofs combine these constructions with the known forward map, finite-moment reconstruction (Kunis et al., 2016), Chebyshev approximation (Trefethen, 2013) and optimal recovery (Ettehad & Foucart, 2020; Foucart, 2021). Appendix I compares the interventions, and Appendices I.1 and I.3 give the equation-level crosswalks.

2 Same data, different partitions

Let logical root r own data D_r and server mass q_r. A factor-K representation partitions it into disjoint children $D_{r1}, \ldots, D_{rK}$, with $D_r = \biguplus_k D_{rk}$ and within-root masses $p_{rk} = |D_{rk}|/|D_r|$. Finite weighted losses are treated in the same way, using each child's share of the total weight. Theorem 1 uses equal-weight examples; the sharp noisy-horizon result allows positive real atomic weights. The root's total server mass remains q_r. Children start from the same round input, take the same number of local steps, and are mass-averaged before the root contributes to the server update.

For one root, write

$$\bar{F}(w) = \sum_{k=1}^{K} p_k F_k(w), \qquad G_F(w) = w - \eta \nabla F(w). \tag{1}$$

With E local steps, the union and partitioned maps are

$$U_E(w) = G_{\bar{F}}^E(w), \qquad P_E(w) = \sum_k p_k G_{F_k}^E(w), \tag{2}$$

and the finite-step repartitioning gap is $D_E(w) = P_E(w) - U_E(w)$.

The primary outcome is $\|D_E(w)\|$; Appendix J specifies the synchronization assumptions and experimental deviations.

The unsplit map $U_E(w)$ fixes the reference for this comparison. The gap measures a change in the update, separately from any change in task utility.

One full-batch gradient step commutes with mass averaging: $P_1(w) = U_1(w)$. More generally, a continuous one-step displacement is invariant under all mass-preserving refinements exactly when

Table 1: Information distinguishes four questions. Each row uses the assumptions of its cited result. The losses stay fixed within each pair; sharpness ranges over the admissible pools, not every prescribed dataset.

Information constraint	What follows	What is not identified
Complete unpaired input derivatives	Sharp two-step envelope $\eta^2 C$ (Theorem 1)	The realized gap
First q endpoints at w_0, actual union spectrum of size q	All later endpoints at w_0 (Theorem 2)	The map at other inputs
Shared matrix spectral weights and first $q-1$ endpoints	Sharp next-gap bound B_q (Theorem 2)	The first unobserved update
Spectral set and noise-bounded common transcript	Uniform diameter $2\mathcal{E}_r$ (Theorem 3)	The radius for a given pool

it is affine in the presented gradient (Proposition 2). A common optimum is a weaker condition. For $F_k(w) = a_k w^2/2$, both P_E and U_E have fixed point zero under $0 < \eta a_k < 1$, but $D_2(w) = \eta^2 \operatorname{Var}_p(a) w$. Appendix A gives the all-horizon comparison with the endpoint-equivalent quadratic surrogate.

3 What unpaired derivatives miss

Let $g_k = \nabla F_k(w)$, $H_k = \nabla^2 F_k(w)$ and use bars for mass-weighted means. For fixed E and sufficiently small η,

$$D_E(w) = \binom{E}{2} \eta^2 \underbrace{\sum_k p_k H_k (g_k - \bar{g})}_{A_{\mathcal{P}}(w)} + O(\eta^3). \tag{3}$$

The coefficient is a signed cross-moment. It can reinforce or cancel across children, so child count alone cannot order the gap.

For partition $\mathcal{P}$, define

$$V_g = \sum_k p_k \|g_k - \bar{g}\|^2, \quad V_H = \sum_k p_k \|H_k - \bar{H}\|_{\mathrm{op}}^2, \quad C_{\mathcal{P}} = \sqrt{V_g V_H}. \tag{4}$$

The centered and pairwise identities give

$$A_{\mathcal{P}} = \frac{1}{2} \sum_{k,\ell} p_k p_\ell (H_k - H_\ell)(g_k - g_\ell), \qquad \|A_{\mathcal{P}}\| \le C_{\mathcal{P}}. \tag{5}$$

For fixed masses, another useful identity is $A_{\mathcal{P}}(w) = \frac{1}{2} \nabla V_g(w)$, obtained by differentiating $V_g = \sum_k p_k \|g_k\|^2 - \|\bar{g}\|^2$. This is familiar Hessian–gradient algebra from gradient-norm modified objectives (Barrett & Dherin, 2021). Unpaired input marginals determine $V_g(w)$ but need not determine this derivative. Section 4 then asks the stronger question of what remains unidentified even when the entire functions V_g and $A_{\mathcal{P}}$ are known.

3.1 Unpaired marginals cannot identify the gap

The same atoms and separate child-level marginals can hide opposite alignments.

Theorem 1 (Unpaired marginals do not identify the gap)**.** *For every $K \ge 3$, $v_g, v_H > 0$, and reference input w_0, there are K^2 equal-weight one-dimensional strictly convex atomic quadratic losses and two equal-mass K-child partitions of those same atoms. At w_0, the partitions have the same union objective and identical multisets $\{g_k(w_0)\}$ and, separately, $\{H_k(w_0)\}$; only their gradient–Hessian pairing differs. Hence they have the same values $V_g(w_0) = v_g$, $V_H(w_0) = v_H$, and $C(w_0) = \sqrt{v_g v_H}$, but $A(w_0) = 0$ for one and $|A(w_0)| = C(w_0)$ for the other. At $E = 2$ their repartitioning gaps from w_0 are exactly 0 and $\eta^2 C(w_0)$. Consequently, no statistic determined only by (V_g, V_H) can identify the leading gap, while C is a sharp worst-case envelope given those marginals.*

For two equal-mass children, swapping the gradient–Hessian pairing changes $A = \frac{1}{4}(H_1 - H_2)(g_1 - g_2)$ to $-A$. Thus its norm is determined by the two unpaired multisets, in any dimension. The zero-versus-positive quadratic two-step norm construction requires $K \geq 3$; this does not identify every two-child signed or nonquadratic map. Figure 1 gives the derivative pairs in a $K = 4$ same-atom realization.

Corollary 1 (Prediction error from unpaired information)**.** *For $E = 2$, any scalar predictor that observes the complete common union objective and the complete unpaired child-gradient and child-Hessian multisets, but not their pairing or A, has worst-case absolute error at least $\eta^2\sqrt{v_g v_H}/2$ on the two partitions in Theorem 1. Moreover, no finite c can make one such prediction satisfy $d/c \leq \widehat{d} \leq cd$ for both gaps d: the zero instance forces $\widehat{d} = 0$, contradicting the positive instance.*

Appendix A.4 gives the general objectives, relates the construction to hierarchical gradient divergences, and records the deterministic $K = 4$ check.

The information loss is not specific to quadratic losses. Adding atom-dependent terms $\lambda c_{ij}(\sin x - x)$, with $x = w - w_0$, leaves both input derivative multisets unchanged but changes the subsequent curvature. For sufficiently small positive steps, the same row and column partitions have distinct two-step endpoints, including a non-cancelling cubic-order correction to their difference. Appendix B gives this strictly convex construction and its exact two-step formula.

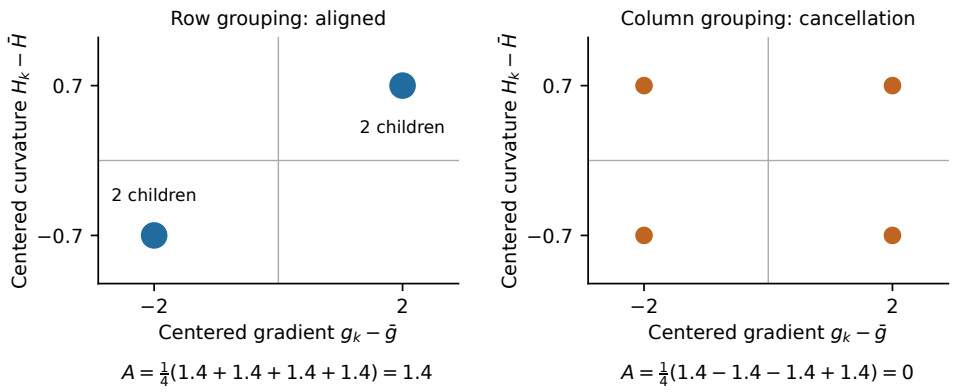

Figure 1: Identical marginals, different pairings. At $w_0 = 0$, the two partitions of sixteen equal-weight losses have the same gradient values and the same curvatures. Row grouping pairs large gradients with large curvatures; column grouping cancels the cross-moment. Here $V_g = 4$, $V_H = .49$, and $C = 1.4$, giving exact two-step gaps .014 and 0 at $\eta = .1$. Appendix A.4 specifies every atom.

4 When early horizons identify a local update

Pairing identifies the two-step quadratic gap. To understand longer horizons, let $\Lambda = \{\lambda_1, \ldots, \lambda_q\}$ be the union of the spectra of the Hessians in two quadratic decompositions. The Hessians are symmetric positive definite, but need not commute. Let $\Pi^b_{k,\lambda}$ project onto the λ-eigenspace of child k in decomposition $b \in \{0, 1\}$, with zero projector when that eigenvalue is absent. Define

$$W^b_\lambda = \sum_k p^b_k \Pi^b_{k,\lambda}, \qquad R^b_\lambda = \sum_k p^b_k \Pi^b_{k,\lambda} g^b_k(w_0).$$

R_λ combines projected child gradients at w_0, allowing cancellation. W_λ retains mass-weighted eigenspace directions; $\mathrm{tr}(W_\lambda)/d$ keeps only spectral mass. Residues determine the reference trajectory; weights determine input dependence. The quadratic map gives

$$P^b_E(w_0; \eta) - w_0 = \sum_{r=0}^{E-1} \binom{E}{r+1} (-\eta)^{r+1} M^b_r, \qquad M^b_r = \sum_{\lambda \in \Lambda} \lambda^r R^b_\lambda.$$

This is a triangular system in vector-valued moments. It contains only powers of each individual Hessian, not products between children.

Theorem 2 (Finite-spectrum identification and sharp disagreement)**.** *At one fixed $\eta_0 \neq 0$, the following conditions are equivalent: agreement of the endpoints at w_0 for horizons $1, \dots, q$; matching $M_0, \dots, M_{q-1}$; and matching R_λ for every $\lambda \in \Lambda$. It identifies every later endpoint at w_0, at every step size. If also $W_\lambda^0 = W_\lambda^1$ for all λ, the maps agree at every input.*
Under this common-spectral-weight condition, suppose $q \geq 2$, $\Lambda \subset [a, b]$, $0 < a < b$, $\Delta = b - a$, and $\|g_k^b(w_0)\| \leq G$. Matching the first $q - 1$ horizons gives

$$\|P_q^1(w;\eta) - P_q^0(w;\eta)\| \leq B_q := \frac{G|\eta|^q \Delta^{q-1}}{2^{2q-4}} \quad \textit{for every } w. \tag{6}$$

The bound is sharp, independently of dimension. Equality is attained by two partitions of the same strictly convex two-dimensional atoms, with noncommuting child Hessians and identical weighted distributions of the full input-gradient vectors.

Vandermonde inversion identifies the vector residues R_λ. The quadratic update is affine in its input:

$$P_E(w) = P_E(w_0) + \left[\sum_{\lambda \in \Lambda} (1 - \eta\lambda)^E W_\lambda\right](w - w_0).$$

For the bound, the monic Chebyshev polynomial of degree $q - 1$ has operator norm at most $\Delta^{q-1}/2^{2q-3}$ on every child Hessian; the lower moments cancel. Appendix F proves the result and constructs the noncommuting equality case. Here q is the number of distinct eigenvalues in the actual union spectrum; constructing a predictor requires supplied support or recurrence information. This is distinct from estimating unknown support from q logs. For noisy logs or approximate support, Appendix H separately bounds observation-error amplification and spectral approximation error, without requiring the actual and nominal spectra to have the same size.

The scalar case has $W_{h_j} = p_j$ and recovers the common weighted curvature-law condition. Its sharp minimax endpoint error is $B_q/2$ on the fixed information class in Appendix E. The compact $q = 3$ noncommuting construction also gives identical entire fields $V_g(w) = G^2 + \Delta^2\|w\|^2/8$ and $A(w) = \Delta^2 w/8$ with different three-step maps. Pairing does not determine higher-order behavior: the equality cases first differ at horizon q, despite agreement at shorter horizons for arbitrary inputs and step sizes.

Why a single test input is not enough. An eigenvalue histogram counts spectral mass but discards its directions. Two five-child decompositions of the same 25 equal-mass two-dimensional losses preserve the objective, the weighted eigenvalue distribution on $\{3, 4, 5\}$, and the full input-gradient distribution. Every child gradient at zero is nonzero, and both decompositions contain noncommuting Hessians. From zero their endpoints agree at every horizon and step size; the shared trajectory is nonzero for $0 < \eta < 1/5$ and $E \geq 1$. Yet

$$P_2^1(w;\eta) - P_2^0(w;\eta) = \tfrac{4}{5}\eta^2 \operatorname{diag}(-1, 1)w.$$

The shared trajectory hides this directional gap; input e_1 exposes it (Figure 2, right). Appendix F specifies every atom. Without shared spectral weights, matching the first q horizons at $d + 1$ affinely independent inputs instead suffices to identify all maps: the test inputs must span the parameter space.

How many noisy observations suffice? Tolerance tests need not estimate q. For quadratic decompositions of a common objective, supply a compact set $S \subset (0, \infty)$ containing all child spectra, a child-input-gradient bound $G > 0$, and bounded recording errors in deterministic endpoint logs. Fix $1 \leq r < L$. At fixed $\eta \neq 0$, put $\phi_m(h) = -\eta \sum_{t=0}^{m-1} (1 - \eta h)^t$ and

$$\mathcal{E}_r = \inf_{b \in \mathbb{R},\, a \in \mathbb{R}^{r-1}} \left\{ G \left\| \phi_L - b - \sum_{m=2}^{r} a_m \phi_m \right\|_{\infty, S} + \sum_{m=2}^{r} |a_m| \varepsilon_m \right\}. \tag{7}$$

The terms bound spectral approximation and amplified recording error, including rounding (Appendix L). The constant b is free because the aggregate gradients agree, so the first horizon adds no constraint on their difference. If the two endpoints at each $m = 2, \dots, r$ share a recorded value within ε_m of each, their target separation is at most $2\mathcal{E}_r$ (Theorem 3). When the following set is nonempty, the smallest depth uniformly guaranteeing tolerance τ is

$$r_\star = \min\{1 \leq r < L : 2\mathcal{E}_r \leq \tau\}.$$

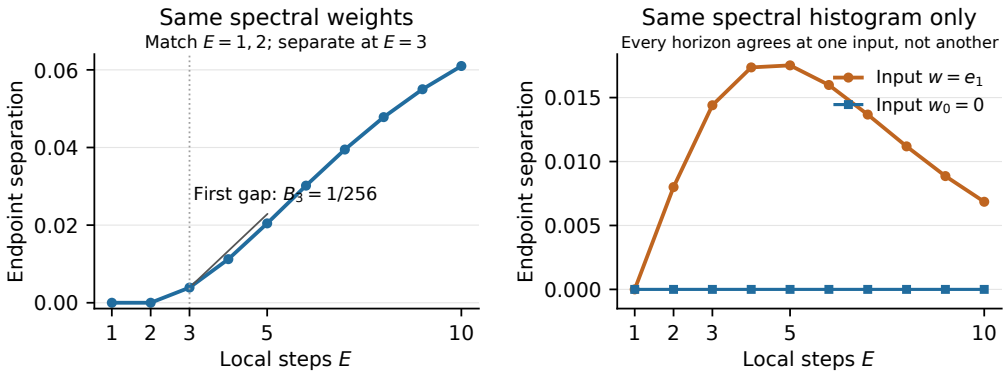

Figure 2: Two distinct identification boundaries, computed from exact quadratic constructions. Left: with common matrix-valued spectral weights and $q = 3$, the two noncommuting partitions agree through two horizons and attain $B_3 = 1/256$ at the third ($G = 1$, spectrum in $[1, 2]$, $\eta = 1/4$). Right: a shared eigenvalue histogram alone permits agreement at every horizon from $w_0 = 0$, yet separation from $w = e_1$ ($\eta = .1$). Both comparisons preserve the atomic losses and objective.

For $S = \{1, 2\}$, $G = 1$, $\eta = 1/4$, $L = 10$, and per-side errors $\varepsilon_m = .001$, the sharp worst-case diameters fall from $2\mathcal{E}_5 \approx .0031918$ to $2\mathcal{E}_6 \approx .0026933$. Tolerance .003 first requires depth six, versus depth two for noiseless identification (Figure 4).

Finite equal-mass realization with atomic bounds. For any integer $N \geq 1$, set $K = 4(N+r+1)$ and $c_N = N/(N+r+1)$. Corollary 4 constructs two equal-mass K-child partitions of one pool of K^2 equal-weight strictly convex two-dimensional quadratic atoms, with $\|P_L^1(w_0) - P_L^0(w_0)\| = 2c_N\mathcal{E}_r$ and a common transcript within the declared error budgets. The pool and its common objective are chosen by the construction. Both partitions share their complete child-Hessian list, matrix spectral weights, and reference-point gradient distribution. Writing $a = \min S$ and $b = \max S$, child spectra remain in S and child gradients at w_0 have norm at most G; atomic gradients there have norm at most $2G$ and atomic Hessian eigenvalues lie in $[a, a + K(b - a)]$. If $b < 2a$, an alternative realization gives the K-independent atomic interval $[2a - b, 2b - a]$.

The construction combines four-angle symmetrization, integer replication with compensated gradient amplitudes, and row/column averaging; its child Hessians are noncommuting when $\mathcal{E}_r > 0$. For a two-point spectrum, four equal-mass children and sixteen equal-weight atoms attain the full diameter $2\mathcal{E}_r$ (Appendix G.4). The ridge study below evaluates these information distinctions on learning instances.

5 Quantifying the Gap Beyond the Pairing Term

The exact two-step construction isolates the missing information. For more general smooth objectives and longer local horizons, the remainder measures what the pairing term alone leaves unresolved.

Proposition 1 (Pairing-based finite-step deviation bounds)**.** *Under the bounded-gradient, bounded-Hessian, Lipschitz-Hessian, and containment conditions in Appendix A.5, let* $\alpha_E = \binom{E}{2}\eta^2$ *and let* $R_{\mathcal{P},E}$ *be the explicit remainder constant there. Then*

$$L_{\mathcal{P}}^A := (\alpha_E\|A_{\mathcal{P}}\| - \eta^3 R_{\mathcal{P},E})_+ \leq \|D_E\| \leq B_{\mathcal{P}}^A := \alpha_E\|A_{\mathcal{P}}\| + \eta^3 R_{\mathcal{P},E} \leq B_{\mathcal{P}}^C := \alpha_E C_{\mathcal{P}} + \eta^3 R_{\mathcal{P},E}. \tag{8}$$

$A_{\mathcal{P}}$ uses K gradients and K Hessian–vector products (Pearlmutter, 1994), without union E-step simulation. A formal interval needs certified child-wise G_k, L_k, ρ_k bounds on one containment region; B^C also needs centered-Hessian operator norms. The looser $C_{\mathcal{P}}$ discards alignment through Cauchy–Schwarz.

$B^C - B^A$ measures discarded pairing information; the $\eta^3 R_{\mathcal{P},E}$ term bounds higher-order dynamics. For candidates i, j, $B_i^A < L_j^A$ guarantees $\|D_i\| < \|D_j\|$. With overlapping intervals, ranking by $\|A\|$ is empirical; evaluate endpoints directly when a definite comparison is needed. Appendix A.10 gives tolerance decisions, inexact-HVP bounds and the refinement identity. For two steps, a cancellation-sensitive refinement replaces the separate endpoint remainders by a bound that vanishes when the initial child gradients agree (Appendix C).

6 Learning examples and prediction boundaries

6.1 A prospective test on logistic objectives

We designate $\|A_{\mathcal{P}}\|$ before evaluating endpoints on 32 FEMNIST writers (Caldas et al., 2018) disjoint from earlier smooth-evaluation cohorts. Each supplies eight hash and eight label-block partitions at $K = 4$. At union-GD states after 25 and 100 steps, we vary $E \in \{2, 5, 10, 20\}$ and $\alpha = \eta L_* \in \{.1, .2, .4, .8\}$, where L_* bounds smoothness for the union and every candidate child. Data, masses and full-batch update rules are matched.

The leading prediction is $Q = \binom{E}{2}\eta^2 A_{\mathcal{P}}$. To distinguish sources of approximation error, we also compute D_{quad}: the exact gap for each loss's quadratic expansion at the input. Thus

$$D - Q = \underbrace{D_{\mathrm{quad}} - Q}_{\text{higher-order composition}} + \underbrace{D - D_{\mathrm{quad}}}_{\text{nonlinear residual}}.$$

At $E = 2$, $D_{\mathrm{quad}} = Q$; at longer horizons, even a quadratic map can depart from the leading prediction. We report mean writer-wise Spearman correlation and the median writer-wise relative vector error. Writer bootstrap intervals and all 64 cells are retained in Appendix O.

Accurate small-step prediction does not persist uniformly. At the fixed small-step comparison, $t = 100$, $E = 2$, and $\alpha = .1$, pairing correlations are .9993 for hash and .9940 for label-block, with median relative vector errors .63% and 1.66%. Subsequently computed $C_{\mathcal{P}}$ and V_H on the same saved inputs provide marginal-curvature comparisons. Pairing improves on both dispersions and their envelope at the two fixed comparisons (Table 17; full grid in Appendix P).

At the stress comparison, $E = 20$, $\alpha = .8$, label-block correlation falls to .4494 (95% interval [.3371,.5543]), while hash retains .8936 ([.8690,.9167]). Relative vector errors are 4.28 and 2.48, respectively (Figure 3). Ordering can remain informative after quantitative prediction fails.

Higher-order composition explains part, but not all, of the loss. At the stress comparison, using the full frozen-quadratic gap improves correlation to .9628 for hash and .6473 for label-block. The nonlinear residual remains substantial: median $\|D - D_{\mathrm{quad}}\|/\|D\|$ is .198 and .349, respectively. The corresponding composition-residual ratios are 2.42 and 3.98. These are norms of separate vectors and do not add as scalar ratios. Of the 64 settings, 28 meet the joint ordering, vector-error and direction criteria; 30 retain correlation at least .7 but exceed .25 relative vector error, and six fall below .7 correlation. Candidate geometry and the complete uncertainty summaries appear in Appendices Q and O.

6.2 Pairing prediction at learned nonlinear states

The pairing contrast persists on 32 writers disjoint from the earlier prediction cohorts, with three fixed initializations of a 409-parameter GELU network. After 20,000 full-batch GD updates with Armijo backtracking, all writers satisfy the fitting and nonlinear-feature-change criteria. The squared-loss local-update comparison uses stateless, fixed-rate GD, $\eta = \alpha/L_{\mathrm{in}}$ (Appendix T).

At $(E, \alpha) = (2, .1)$, pairing predicts both ordering and vectors (Table 2). In label-block partitions, $C_{\mathcal{P}}$ and V_g rank negatively for every seed while $\|A_{\mathcal{P}}\|$ remains predictive. The envelope $C_{\mathcal{P}}$ need not rank gaps well. Here $D_{\mathrm{quad}} = Q$ is an identity. All six seed–family panels also meet the joint prediction criteria at the fixed secondary $(5, .1)$ comparison: median vector errors are .1151–.2008, with smaller frozen-quadratic errors in each panel.

Only 27/96 settings meet the joint criteria; all six $(20, .8)$ stress panels fail (vector errors 5.674–8.117). The affine reference fits better on every writer. Appendix S retains the earlier constant-rate study: late

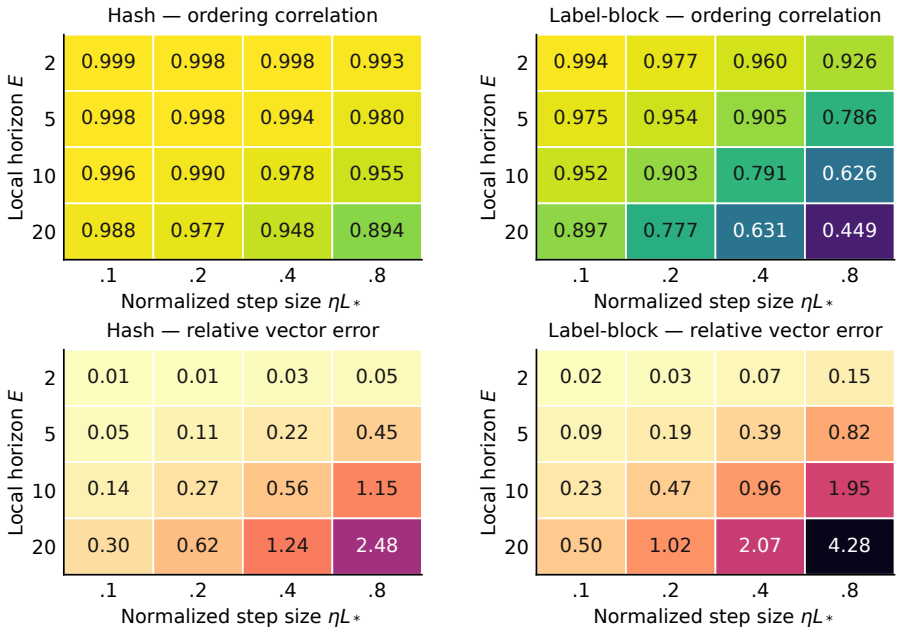

Figure 3: Logistic prediction across the full horizon–step-size grid at the 100-step input. Top: mean writer-wise Spearman correlation between $\|A\|$ and the exact gap. Bottom: median of writer-wise candidate-median vector errors $\|D - Q\|/\|D\|$. Each entry uses the same 32 writers and eight membership-distinct candidates per family. The 25-step grid is reported in Appendix O. Magnitude accuracy degrades before ordering, especially for hash partitions.

Table 2: Independent-cohort prediction at $(E, \alpha) = (2, .1)$. Correlations are means over 32 writers; error and cosine are medians of writer medians. Seeds 1–3 are fixed initializations on the same writers, not independent cohorts. Appendix T gives pointwise writer-bootstrap intervals; the numerical supplement contains all 96 settings.

Seed	Partition	ρ_A	ρ_C	ρ_{V_g}	Rel. error	Cosine
1	Hash	0.9955	0.8743	0.8185	0.0346	0.9995
1	Label-block	0.9658	-0.2530	-0.1503	0.0423	0.9992
2	Hash	0.9970	0.8713	0.8080	0.0379	0.9995
2	Label-block	0.9643	-0.2939	-0.1935	0.0454	0.9992
3	Hash	0.9918	0.8185	0.7708	0.0479	0.9991
3	Label-block	0.9643	-0.4182	-0.2582	0.0431	0.9994

oscillations, 5/32 supported settings at step 20,000 and 8/32 at step 2,000. Different cohorts prevent attributing their difference to backtracking. Appendix U summarizes the earlier near-initialization controls.

6.3 What can be inferred from three ridge logs?

Three-log matrix recovery resolves the primary ridge tests; matched scalar-block recovery and restricted-spectrum bounds abstain (Table 3). We use eight disjoint CIFAR-100 subsets (Krizhevsky, 2009), 64-dimensional frozen ResNet-18 features (He et al., 2016) and eight equal children. The original and matched readouts receive covariances, the fitted head and logs at horizons 1–3, but no raw labels or later endpoints, to predict $L \in \{4, 6, 10\}$. We measure $s_L = \|C_{\text{probe}}^{1/2} D_L\|_F$, the RMS change of the full class-score vector across unlabeled probe rows, where $C_{\text{probe}} = n^{-1} \sum_i x_i x_i^\top$. All eight primary score changes, at $(\alpha, L, \tau) = (.75, 6, .01)$, are below tolerance (median 7.12×10^{-4}).

Table 3: Three comparisons on the original eight pools. Entries count correct primary passes from three logs; unresolved cases abstain.

Comparison	Available information and constraints	Primary result
Original readouts	Shared covariances, fitted head, fit budget and logs. Matrix recovery also uses joint fit/log constraints and child-wise label-feasibility ellipsoids.	Matrix: 8/8; scalar interpolation: 0/8.
Matched families	Same feasible set, fit/log constraints, normalization, regularized objective and error accounting. Recovery blocks $Q_m = a_m I_d$ versus $Q_m \in \mathbb{R}^{d\times d}$.	Matrix: 8/8; scalar-block: 0/8.
Restricted spectrum	Spectral set, input-gradient bound, early-discrepancy budgets and probe norm; no full matrices or directional log arrays.	Interval or full numerical support: 0/8.

Matched recovery. (i) Form shared fit/log constraints and child-feasibility bounds. (ii) Fit scalar or matrix recovery blocks with the same normalization and regularized objective; in original units predict $\widehat{D}_L = d_L - Rz$. (iii) Bound the operator residual and amplified errors; intersect intervals within each family. Pass if the upper endpoint is $\leq \tau$, reject if the lower is $> \tau$, otherwise abstain (Appendix L.4). The original comparison changes constraint use; the matched secondary analysis isolates a coefficient-family benefit for its specified objective, not optimality over all scalar estimators. Zero/one-log controls resolve none. Table 9 and Appendix L retain smaller-step scalar successes, fit and rounding terms. With budgets from the same early observations, the restricted theorem gives three-log interval-spectrum bounds .1767–.1803. Full numerical support tightens these without resolving .01. Both spectral sets resolve all eight at $\alpha = .25$, $L = 4, 6$, but not $L = 10$ (Appendix V).

Prediction on unused pools. On eight further disjoint subsets, we reuse features and probes and fix predictions before later endpoints. At $\alpha = .75$, $L = 6$, $\tau = .01$, three-log matrix recovery passes all eight fitted-head and all eight zero-head tests: paired states of eight pools, with every target below tolerance. Scalar-block recovery, zero/one-log controls and restricted-spectrum bounds resolve none. At the prespecified secondary tolerance .001, the matrix readout passes seven fitted-head cases and abstains in one; on zero heads it passes seven and correctly rejects one. Appendix W reports all cases and tolerances.

Agreement on a probe. On the original pools, matrix forecasts establish 99% union classification agreement in eight fitted-head cases at $L = 4, 6$, not $L = 10$. A cold-start replay instead has 1.48–4.22% disagreement at $L = 6$: three logs exclude 99% agreement in seven cases and abstain in one. The scalar baseline and one-log control remain inconclusive; all bounds cover their targets (Appendices M and N).

Supporting bounds and controls. Appendix K.1 reports Float64 coverage and tightness at initialization and nonzero states. Appendix U summarizes earlier ranking, batching, activation and adapter controls; separated endpoints accompany mixed downstream effects.

7 Discussion and conclusion

Matching objectives, reference trajectories, and whole update maps are distinct requirements. Pairing controls two-step ambiguity; curvature-weighted moments and matrix spectral weights govern longer quadratic updates and input dependence. Our same-loss constructions establish sharpness, including for noncommuting Hessians.

Smooth guarantees require deterministic full-batch GD and derivative bounds on a common containment region; sampling, persistent state, BatchNorm and nonsmooth updates need separate analyses. Dense logistic B^C costs more than simulation (Appendix U.2); probe bounds concern agreement, not population accuracy. Prediction tests use FEMNIST logistic and small GELU models at fixed trained states; adapter and task results establish separation, not quality gains.

ETHICS STATEMENT

We use public benchmarks, collect no new human-subject data, and do not attempt re-identification. A shard is not a new legal, privacy, or consent unit; governance and privacy accounting should remain tied to the logical owner. Label- and source-order partitions are stress tests, not deployment advice. Gradients, Hessian–vector products, and metadata can leak information and should remain within the data owner's authorized access boundary; the endpoint bounds provide no privacy mechanism.

REPRODUCIBILITY STATEMENT

The appendices provide proofs, constructions, protocols, models, comparison rules, costs and limitations. The numerical supplement retains complete fixed-grid tables, including unsuccessful settings, and small deterministic checks of the mathematical examples. The full training and orchestration code, checkpoints and raw execution records are not included in this submission.

Reading guide. The main text states the identification results and their empirical scope. For proofs, see Appendices A, F and G; for the recovery protocol, see Appendix L. Appendix U collects the applicability boundaries and unsuccessful controls. Complete setting-level tables are provided separately for optional numerical inspection.

AI USE STATEMENT

Generative AI assisted with reviewing experimental protocols and code, checking mathematical arguments, interpreting results, searching the literature, preparing figures and tables, and drafting and editing the manuscript. AI-assisted content was checked against source materials, experimental results, and verification tests, as applicable. Synthetic loss constructions are specified analytically and checked numerically; benchmark examples were not generated by AI. No qualitative human-response analysis was performed. The authors take responsibility for the final content.

REFERENCES

David G. T. Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3q5IqUrkcF.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018. URL https://arxiv.org/abs/1812.01097.

Timothy Castiglia, Anirban Das, and Stacy Patterson. Multi-level local SGD: Distributed SGD for heterogeneous hierarchical networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=C70cp4Cn32.

Zachary Charles and Jakub Konečný. Convergence and accuracy trade-offs in federated learning and meta-learning. In *AISTATS*, 2021.

Zachary Charles and Keith Rush. Iterated vector fields and conservatism, with applications to federated learning. In *Algorithmic Learning Theory*, pp. 130–147, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Mahmood Ettehad and Simon Foucart. Instances of computational optimal recovery: Dealing with observation errors. *arXiv preprint arXiv:2004.00192*, 2020. URL https://arxiv.org/abs/2004.00192.

Simon Foucart. Instances of computational optimal recovery: Refined approximability models. *Journal of Complexity*, 62:101503, 2021. doi: 10.1016/j.jco.2020.101503. URL https://foucart.github.io/publi/OR_Overparametrization_v2.pdf.

Walter Gautschi. How (un)stable are Vandermonde systems? In Roderick Wong (ed.), *Asymptotic and Computational Analysis*, volume 124 of *Lecture Notes in Pure and Applied Mathematics*, pp. 193–210. Marcel Dekker, 1990. URL https://www.cs.purdue.edu/homes/wxg/selected_works/section_01/118.pdf.

Margalit R. Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (Local SGD) and continuous perspective. In *AISTATS*, pp. 9050–9090, 2022.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: Unified theory and new efficient methods. In *AISTATS*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck R. Cadambe. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11082–11094, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/c17028c9b6e0c5deaad29665d582284a-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Yan Huang, Ying Sun, Zehan Zhu, Changzhi Yan, and Jinming Xu. Tackling data heterogeneity: A new unified framework for decentralized SGD with sample-induced topology. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 9310–9345, 2022. URL https://proceedings.mlr.press/v162/huang22i.html.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

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

Stefan Kunis, Thomas Peter, Tim Römer, and Ulrich von der Ohe. A multivariate generalization of Prony's method. *Linear Algebra and its Applications*, 490:31–47, 2016. doi: 10.1016/j.laa.2015.10.023. URL https://arxiv.org/abs/1506.00450.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020a. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2020b. URL https://arxiv.org/abs/1907.02189.

Ruichen Luo, Sebastian U. Stich, Samuel Horváth, and Martin Takáč. Revisiting LocalSGD and SCAFFOLD: Improved rates and missing analysis. In *AISTATS*, pp. 2539–2547, 2025.

Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtárik. From local SGD to local fixed-point methods for federated learning. In *ICML*, 2020.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

NIST. Digital library of mathematical functions, section 18.38(i): Mathematical applications, 2026a. URL https://dlmf.nist.gov/18.38. Online reference, accessed September 9, 2026.

NIST. Digital library of mathematical functions, section 3.3: Interpolation, 2026b. URL https://dlmf.nist.gov/3.3. Online reference, accessed September 9, 2026.

Kumar Kshitij Patel, Ali Zindari, Sebastian Stich, and Lingxiao Wang. Revisiting consensus error: A fine-grained analysis of local SGD under second-order data heterogeneity. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/54cee855c7eb9429b4b3668d06bf670f-Abstract-Conference.html.

Kumar Kshitij Patel, Rustem Islamov, Sebastian U. Stich, Aurelien Lucchi, Eduard Gorbunov, and Lingxiao Wang. What's in a smoothness constant? tighter rates for local SGD with bounded second-order heterogeneity. *arXiv preprint arXiv:2607.14731*, 2026. URL https://arxiv.org/abs/2607.14731.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.09767.

Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. Society for Industrial and Applied Mathematics, 2013. URL https://people.maths.ox.ac.uk/trefethen/ATAP/.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *NeurIPS*, 2020.

Jianyu Wang, Zheng Xu, Zachary Garrett, Zachary Charles, Luyang Liu, and Gauri Joshi. Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*, 2021. Presented at FL-ICML'21.

Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. Demystifying why local aggregation helps: Convergence analysis of hierarchical SGD. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8548–8556, 2022. doi: 10.1609/aaai.v36i8.20832. URL https://ojs.aaai.org/index.php/AAAI/article/view/20832.

Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. A new theoretical perspective on data heterogeneity in federated optimization. In *ICML*, pp. 51650–51700, 2024.

Blake E. Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6281–6292, 2020. URL https://arxiv.org/abs/2006.04735.

Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision*, pp. 3–19, 2018. URL https://arxiv.org/abs/1803.08494.

A Proofs and deterministic checks

A.1 One-step characterization and the multi-step term

Proposition 2 (Fixed-state one-step characterization)**.** *Fix a round input w and optimizer state s, and write the one-step endpoint presented with gradient g as $w + \Delta_{w,s}(g)$, where $\Delta_{w,s} : \mathbb{R}^d \to \mathbb{R}^d$ is continuous. For every finite collection of gradients $g_1, \dots, g_K$ and masses $p_k \geq 0$ with $\sum_k p_k = 1$, universal mass-preserving one-step invariance means*

$$w + \Delta_{w,s}\left(\sum_k p_k g_k\right) = \sum_k p_k \big(w + \Delta_{w,s}(g_k)\big). \tag{9}$$

It holds if and only if there exist $M_{w,s} \in \mathbb{R}^{d\times d}$ and $b_{w,s} \in \mathbb{R}^d$ such that

$$\Delta_{w,s}(g) = M_{w,s} g + b_{w,s}. \tag{10}$$

If $\Delta_{w,s}(0) = 0$, then $b_{w,s} = 0$.

Proof. After cancelling w, Eq. 9 is Jensen's affine identity. Applying it first to rational convex mixtures and then using continuity proves necessity of $\Delta_{w,s}(g) = M_{w,s} g + b_{w,s}$; direct substitution proves sufficiency. Evaluating at $g = 0$ gives the last statement. □

The proposition is fixed-state: independently evolving persistent optimizer states are outside its scope.

For the gradient step $\Delta(g) = -\eta g$, exact one-step invariance follows: $\sum_k p_k(w - \eta g_k) = w - \eta \sum_k p_k g_k$. For a smooth objective F, write $g_F = \nabla F(w)$ and $H_F = \nabla^2 F(w)$. For fixed E, induction gives

$$G_F^E(w) = w - E\eta g_F + \binom{E}{2}\eta^2 H_F g_F + O(\eta^3). \tag{11}$$

The claim is immediate at $E = 1$. If it holds at step t, Taylor expansion around w gives $\nabla F(w_t) = g_F - t\eta H_F g_F + O(\eta^2)$. Substitution into $w_{t+1} = w_t - \eta \nabla F(w_t)$ changes the second-order coefficient from $\binom{t}{2} + t$ to $\binom{t+1}{2}$. Applying the expansion to every child and to $\bar{F}$ cancels first-order terms and gives

$$\begin{aligned} P_E(w) - U_E(w) &= \binom{E}{2}\eta^2 \left(\sum_k p_k H_k g_k - \bar{H}\bar{g}\right) + O(\eta^3) \\ &= \binom{E}{2}\eta^2 \sum_k p_k H_k (g_k - \bar{g}) + O(\eta^3). \end{aligned} \tag{12}$$

A.2 PARTITION-SPECIFIC COUPLING

Because $\sum_k p_k(H_k - \bar{H}) = 0$ and $\sum_k p_k(g_k - \bar{g}) = 0$,

$$\begin{aligned} A_{\mathcal{P}}(w) &= \sum_k p_k H_k(g_k - \bar{g}) \\ &= \sum_k p_k(H_k - \bar{H})(g_k - \bar{g}) \\ &= \frac{1}{2}\sum_{k,\ell} p_k p_\ell (H_k - H_\ell)(g_k - g_\ell). \end{aligned} \tag{13}$$

Expanding the double sum proves the last equality. Weighted Cauchy–Schwarz yields

$$\begin{aligned} \|A_{\mathcal{P}}(w)\| &\le \sum_k p_k \|H_k - \bar{H}\|_{\mathrm{op}} \|g_k - \bar{g}\| \\ &\le \left(\sum_k p_k \|H_k - \bar{H}\|_{\mathrm{op}}^2\right)^{1/2} \left(\sum_k p_k \|g_k - \bar{g}\|^2\right)^{1/2} = C_{\mathcal{P}}(w). \end{aligned} \tag{14}$$

Combining this inequality with Eq. 30 proves the scalar upper certificate. At server level, for root masses q_r,

$$\left\| \sum_r q_r D_r \right\| \le \sum_r q_r \mathcal{B}_{\mathcal{P}_r, E_r}(w). \tag{15}$$

The certificate is sufficient, not necessary: root-level vectors may cancel, and operator-norm dispersion may be conservative.

A.3 REFINEMENT CALCULUS AND INEXACT COUPLING

Let parent block b have mass q_b and conditional child masses $p_{k|b}$. Expanding the fine-partition cross-moment around the parent means gives the weighted law of total covariance,

$$\sum_{b,k} q_b p_{k|b}(H_k - \bar{H})(g_k - \bar{g}) = \sum_b q_b (H_b - \bar{H})(g_b - \bar{g}) \tag{16}$$

$$+ \sum_b q_b \sum_{k \in b} p_{k|b}(H_k - H_b)(g_k - g_b). \tag{17}$$

All mixed terms vanish because the conditional centered gradients and Hessians sum to zero. This proves Eq. 38.

For a two-child block, multiplying its conditional coupling by $p_a + p_b$ yields

$$\Delta_{ab} = \frac{p_a p_b}{p_a + p_b}(H_a - H_b)(g_a - g_b). \tag{18}$$

After merging, $A_{\mathrm{new}} = A_{\mathrm{old}} - \Delta_{ab}$, and expanding $\|A_{\mathrm{old}} - \Delta_{ab}\|^2 < \|A_{\mathrm{old}}\|^2$ proves the stated merge criterion. Finally, $\|A\| \in [(\|\widetilde{A}\| - \zeta)_+, \|\widetilde{A}\| + \zeta]$ by the triangle inequality; substituting these endpoints into Eq. 30 proves Eq. 37.

A.4 MARGINAL-DISPERSION NONIDENTIFIABILITY

Translate coordinates so the common reference input is $w_0 = 0$. For any $K \ge 3$, let $r = -2 + \sqrt{3}$ and define

$$s_0 = (1, r, -1 - r, 0, \ldots, 0), \qquad t_0 = (r, 1, -1 - r, 0, \ldots, 0). \tag{19}$$

Set $s = \sqrt{K/\|s_0\|^2}\, s_0$ and $t = \sqrt{K/\|t_0\|^2}\, t_0$. Since $r^2 + 4r + 1 = 0$ and t_0 is a permutation of s_0, these vectors satisfy

$$\mathbf{1}^\top s = \mathbf{1}^\top t = s^\top t = 0, \qquad \|s\|^2 = \|t\|^2 = K, \qquad \{s_i\}_{i=1}^K = \{t_i\}_{i=1}^K. \tag{20}$$

Fix arbitrary $v_g, v_H > 0$, choose any mean gradient $\bar{g}$, set $M = \max_{i,j} |s_i + t_j|$, and choose $\bar{H} > M\sqrt{v_H}$. For $i, j \in \{1, \dots, K\}$, define K^2 equally weighted atomic losses

$$\ell_{ij}(x) = \frac{1}{2}\left[\bar{H} + \sqrt{v_H}(s_i + t_j)\right] x^2 + \left[\bar{g} + \sqrt{v_g}(s_i + s_j)\right] x. \tag{21}$$

Every atom is strictly convex because its curvature is at least $\bar{H} - M\sqrt{v_H} > 0$.

Grouping the same atoms by rows gives

$$g_i^{\text{row}} = \bar{g} + \sqrt{v_g} s_i, \qquad H_i^{\text{row}} = \bar{H} + \sqrt{v_H} s_i, \tag{22}$$

whereas grouping them by columns gives

$$g_j^{\text{col}} = \bar{g} + \sqrt{v_g} s_j, \qquad H_j^{\text{col}} = \bar{H} + \sqrt{v_H} t_j. \tag{23}$$

Both are equal-mass K-child partitions of exactly the same K^2 atoms and induce the union objective $\bar{F}(x) = \bar{H}x^2/2 + \bar{g}x$. At w_0, both have the same child-gradient multiset because both use the entries of s. Their child-Hessian multisets are also the same because t is a permutation of s; only the pairing of gradients and Hessians changes. Both therefore have $V_g = v_g$, $V_H = v_H$, and $C = \sqrt{v_g v_H}$, while

$$A^{\text{row}} = \tfrac{1}{K} \sum_i (\sqrt{v_H} s_i)(\sqrt{v_g} s_i) = C, \tag{24}$$

$$A^{\text{col}} = \tfrac{1}{K} \sum_j (\sqrt{v_H} t_j)(\sqrt{v_g} s_j) = 0 \tag{25}$$

because $s^\top t = 0$. The marginal equalities are asserted at w_0; they need not remain equal away from that input.

Two gradient steps on a quadratic child from zero equal $-2\eta g_k + \eta^2 H_k g_k$. Thus $D_2 = \eta^2 A$ exactly, so the row and column partitions have gaps $\eta^2 C$ and zero. Equality for the row partition also shows that C is a sharp worst-case envelope given the two marginals.

The same construction directly compares with the gradient-divergence summaries used for hierarchical grouping. Treat the K^2 atoms as workers and the rows or columns as K equal groups. At w_0, the group-to-union (upward) dispersion is v_g for either partition, and the worker-to-group (downward) dispersion is also v_g for either partition: within a row the residual is $\sqrt{v_g} s_j$, while within a column it is $\sqrt{v_g} s_i$. These equalities hold at the reference input; they do not assert the uniform-in-input divergence conditions of an H-SGD convergence theorem.

The deterministic numerical check uses $K = 4$, $s = (1, -1, 1, -1)$, $t = (1, 1, -1, -1)$, and the same sixteen atomic objectives. It obtains $V_g = 4$, $V_H = 0.49$, $C = A^{\text{row}} = 1.4$, $A^{\text{col}} = 0$, and two-step gaps 0.014 and zero algebraically (column float64 residual 3.47×10^{-17}). For this construction, the row and column child-gradient multisets are exactly the same: each contains two copies of $\bar{g} + \sqrt{v_g}$ and two copies of $\bar{g} - \sqrt{v_g}$. Their child-Hessian multisets are also exactly the same: each contains two copies of $\bar{H} + \sqrt{v_H}$ and two copies of $\bar{H} - \sqrt{v_H}$. Thus the two partitions differ in the pairing between gradient and curvature, not in either child-level marginal multiset.

A.5 Finite-step remainder certificate

For a nonnegative step size η, let a common convex set $\mathcal{C}$ contain w, every child and union iterate through step E, and the line segments from w to those iterates. Throughout $\mathcal{C}$, suppose

$$\|\nabla F_k(x)\| \le G_k, \quad \|\nabla^2 F_k(x)\| \le L_k, \quad \|\nabla^2 F_k(x) - \nabla^2 F_k(y)\| \le \rho_k \|x - y\|. \tag{26}$$

Set $\bar{G} = \sum_k p_k G_k$, $\bar{L} = \sum_k p_k L_k$, $\bar{\rho} = \sum_k p_k \rho_k$, and, with $\binom{E}{3} = 0$ for $E < 3$, define

$$B_E(G, L, \rho) := \binom{E}{3} L^2 G + \frac{E(E-1)(2E-1)}{12} \rho G^2. \tag{27}$$

Then

$$D_E(w) = \binom{E}{2} \eta^2 A(w) + r_E, \tag{28}$$

$$\|r_E\| \le \eta^3 R_{\mathcal{P},E}, \qquad R_{\mathcal{P},E} := \sum_k p_k B_E(G_k, L_k, \rho_k) + B_E(\bar{G}, \bar{L}, \bar{\rho}). \tag{29}$$

Consequently,

$$\left(\binom{E}{2}\eta^2\|A(w)\| - \eta^3 R_{\mathcal{P},E}\right)_+ \le \|D_E(w)\| \le \binom{E}{2}\eta^2\|A(w)\| + \eta^3 R_{\mathcal{P},E}. \tag{30}$$

For the proof, consider one objective and write $z_t = G_F^t(w)$, $\delta_t = z_t - w$, $g = \nabla F(w)$, and $H = \nabla^2 F(w)$. The trajectory bound is $\|\delta_t\| \le t\eta G$. Hessian Lipschitzness gives

$$\nabla F(w+\delta_t) = g + H\delta_t + q_t, \qquad \|q_t\| \le \tfrac{1}{2}\rho\|\delta_t\|^2. \tag{31}$$

For $h_t := \delta_t + t\eta g$, telescoping the full gradients gives

$$h_t = -\eta\sum_{s=0}^{t-1}[\nabla F(w+\delta_s) - g].$$

The Hessian bound along each containing segment implies $\|\nabla F(w+\delta_s) - g\| \le L\|\delta_s\|$, so

$$\|h_t\| \le \eta L\sum_{s=0}^{t-1}\|\delta_s\| \le \binom{t}{2}\eta^2 LG. \tag{32}$$

Summing the updates and substituting $\delta_t = -t\eta g + h_t$ gives

$$z_E = w - E\eta g + \binom{E}{2}\eta^2 Hg + e_E, \qquad e_E = -\eta H\sum_{t=0}^{E-1} h_t - \eta\sum_{t=0}^{E-1} q_t. \tag{33}$$

Using $\sum_{t=0}^{E-1}\binom{t}{2} = \binom{E}{3}$ and $\sum_{t=0}^{E-1} t^2 = E(E-1)(2E-1)/6$ gives $\|e_E\| \le \eta^3 B_E(G, L, \rho)$. Apply this expansion to each child and to $\bar{F}$, then use linearity and the triangle inequality.

A.6 Task-linked tolerances

Let h be an M-Lipschitz scalar evaluation functional on a set containing the two endpoints. Then $B_{\mathcal{P}}^A \le \varepsilon$ implies

$$|h(P_E(w)) - h(U_E(w))| \le M\varepsilon. \tag{34}$$

The same implication holds with B^C. Thus a task-level tolerance τ can induce $\varepsilon = \tau/M$ when a valid M is available. The transfer requires a valid Lipschitz constant for the chosen task quantity and does not make the union endpoint task-optimal.

A.7 Finite-horizon round propagation

Let P_t and U_t be the partitioned and union round maps and define their same-input defect $d_t(x) := P_t(x) - U_t(x)$. If P_t is α_t-Lipschitz on the relevant domain, then, for $p_{t+1} = P_t(p_t)$, $u_{t+1} = U_t(u_t)$, and $e_t = p_t - u_t$,

$$\|e_T\| \le \|e_0\|\prod_{s=0}^{T-1}\alpha_s + \sum_{t=0}^{T-1}\|d_t(u_t)\|\prod_{s=t+1}^{T-1}\alpha_s. \tag{35}$$

The identity $e_{t+1} = P_t(p_t) - P_t(u_t) + d_t(u_t)$ gives the one-step recurrence; induction proves the result. Signed vector defects may cancel, so the bound is not a directional or convergence claim.

A.8 Fixed-point agreement does not imply finite-step agreement

Let $F_k(w) = a_k w^2/2$, $p_k > 0$, $\sum_k p_k = 1$, and $r_k = 1 - \eta a_k \in (0,1)$. Write $\bar{r} = \sum_k p_k r_k$ and $s_E = \sum_k p_k r_k^E$. Direct iteration gives $U_E(w) = \bar{r}^E w$ and $P_E(w) = s_E w$. Both maps have unique fixed point zero, also the minimizer of $\bar{F}$ and of the exact endpoint-equivalent surrogate $\widetilde{F}_E(w) = (1-s_E)w^2/(2\eta)$. Nevertheless, for integer $E \ge 2$, heterogeneous curvatures, and $w \ne 0$, strict Jensen convexity gives

$$0 < \frac{D_E(w)}{w} = s_E - \bar{r}^E,$$
$$\frac{E(E-1)}{2} r_{\min}^{E-2}\eta^2\mathrm{Var}_p(a) \le s_E - \bar{r}^E \le \frac{E(E-1)}{2} r_{\max}^{E-2}\eta^2\mathrm{Var}_p(a). \tag{36}$$

At $E = 2$, $D_2(w) = \eta^2\mathrm{Var}_p(a)w$ exactly.

A.9 Deterministic numerical verification

Numerical verification samples four seeded positive-definite quadratic child objectives in five dimensions. The maximum one-step error over $\eta \in \{0.2, 0.1, 0.05, 0.025\}$ is 4.45×10^{-16}; at five local steps, the residual after the second-order term has fitted order 2.9398 over the three smallest learning rates. The shared-atom check uses $v_g = 4$, $v_H = 0.49$, $\bar{g} = 0.3$, $\bar{H} = 2$, and $\eta = 0.1$. It verifies all sixteen shared atoms, both complete partitions, minimum atomic curvature 0.6, $A^{\mathrm{row}} = C = 1.4$, $A^{\mathrm{col}} = 0$, row gap 0.014, and column float64 residual 3.47×10^{-17}.

A.10 Consequences of the finite-step bounds

The bounds in Proposition 1 also give the following tolerance and ordering consequences. They are auxiliary to the identification result; no runtime or partition-optimization claim is needed.

If an approximate HVP computation supplies $\widetilde{A}$ together with a verified deterministic error budget $\|\widetilde{A} - A\| \le \zeta$, the same argument gives the valid inexact interval

$$\Big(\alpha_E(\|\widetilde{A}\| - \zeta)_+ - \eta^3 R_{\mathcal{P},E}\Big)_+ \le \|D_E\| \le \alpha_E(\|\widetilde{A}\| + \zeta) + \eta^3 R_{\mathcal{P},E}. \tag{37}$$

Equation 37 applies to an approximate coupling estimate accompanied by a verified deterministic error bound ζ.

Corollary 2 (Ternary decision and certified ranking)**.** *For a same-input parameter tolerance ε, $B^A_{\mathcal{P}} \le \varepsilon$ certifies ε-equivalence to the union endpoint, whereas $L^A_{\mathcal{P}} > \varepsilon$ certifies ε-separation; otherwise the certificate abstains. For candidates i, j, if $B^A_i < L^A_j$, then $\|D_i\| < \|D_j\|$.*

The looser condition $B^C_{\mathcal{P}} \le \varepsilon$ also certifies equivalence using only the marginal envelope. An upper bound above tolerance, whether $B^A > \varepsilon$ or $B^C > \varepsilon$, does not by itself establish separation. When intervals overlap, ranking by $A_{\mathcal{P}}$ or $C_{\mathcal{P}}$ is an empirical heuristic whose accuracy must be measured.

A.11 How refinement changes the coupling

Refinement adds signed within-block terms, so adding children can either raise or lower the leading gap.

Proposition 3 (Refinement calculus and an exact leading-coupling merge rule)**.** *If a fine partition $\mathcal{Q}$ refines parent blocks b of $\mathcal{P}$, then, at the common input,*

$$A_{\mathcal{Q}} = A_{\mathcal{P}} + \sum_b q_b \sum_{k \in b} p_{k|b}(H_k - H_b)(g_k - g_b). \tag{38}$$

Thus child count alone has no monotone implication: block contributions can align or cancel. Merging two current children a, b changes the coupling from A to $A - \Delta_{ab}$, where

$$\Delta_{ab} = \frac{p_a p_b}{p_a + p_b}(H_a - H_b)(g_a - g_b). \tag{39}$$

The merge strictly reduces the leading norm iff $2\langle A, \Delta_{ab}\rangle > \|\Delta_{ab}\|^2$. After the child gradients and HVPs are available, an exact best single merge for the leading coupling is therefore any $(a, b)^* \in \arg\min_{a<b} \|A - \Delta_{ab}\|$.

Evaluating every pair exactly requires either stored Hessians or pair-specific HVPs. The rule characterizes the best single merge for the leading coupling; its cost depends on how those derivatives are obtained.

B A nonquadratic matched construction

A smooth nonquadratic perturbation preserves the construction's exact input marginals while introducing a higher-order difference between its two partitioned maps. The following scalar example is engineered for this comparison; it is separate from the observational FEMNIST ranking analysis.

Shared atoms and matched input derivatives. For $K \in \{3, 4, 8\}$, let

$$r = -2 + \sqrt{3}, \qquad s^0 = (1, r, -1 - r, 0, \ldots, 0), \qquad s = \frac{\sqrt{K}}{\|s^0\|} s^0,$$

and obtain t by interchanging the first two coordinates of s. Then $\mathbf{1}^\top s = \mathbf{1}^\top t = s^\top t = 0$, $\|s\|^2 = \|t\|^2 = K$, and s, t are permutations of one another. Set $w_0 = 1$, $x = w - w_0$, $M = \max_{i,j} |s_i + t_j|$, and $\bar{h} = 1 + 0.7M$. Define K^2 equal-weight atomic losses by

$$f_{ij}(w) = g_{ij} x + \tfrac{1}{2} h_{ij} x^2 + \lambda c_{ij} (\sin x - x), \tag{40}$$
$$g_{ij} = 0.3 + 2(s_i + s_j), \qquad h_{ij} = \bar{h} + 0.7(s_i + t_j).$$

The quadratic control uses $\lambda = 0, c_{ij} = 1$. Two nonquadratic controls use $\lambda = 1/2$: the common control has $c_{ij} = 1$, whereas the varying control has

$$c_{ij} = 1 + \frac{s_i + t_j}{2M}.$$

Within each control, row and column grouping partition exactly the same atoms into K equal-mass children. Their common union objective is

$$\bar{F}(w) = 0.3x + \tfrac{1}{2} \bar{h} x^2 + \lambda(\sin x - x).$$

All atoms are globally strictly convex: $h_{ij} \geq 1$ and $|c_{ij}| \leq 3/2$ give

$$f''_{ij}(w) = h_{ij} - \lambda c_{ij} \sin x \geq 1 - \tfrac{1}{2}\tfrac{3}{2} = \tfrac{1}{4} > 0.$$

The perturbation's first two derivatives vanish at $x = 0$. Consequently, at w_0 the row and column child derivatives are exactly

$$g_i^{\mathrm{row}} = g_i^{\mathrm{col}} = 0.3 + 2s_i, \qquad h_i^{\mathrm{row}} = \bar{h} + 0.7 s_i, \qquad h_i^{\mathrm{col}} = \bar{h} + 0.7 t_i.$$

Both complete separate gradient and Hessian multisets therefore match, with $V_g = 4$, $V_H = 0.49$ and $C = 1.4$, while $A_{\mathrm{row}} = 1.4$ and $A_{\mathrm{col}} = 0$. In the varying control the child perturbation coefficients are $c_i^{\mathrm{row}} = 1 + s_i/(2M)$ and $c_i^{\mathrm{col}} = 1 + t_i/(2M)$; in the common control they all equal one.

Exact two-step correction. For a child with coefficients (g, h, c), two GD steps of size $\eta \geq 0$ from w_0 have centered endpoint

$$x_2(g, h, c) = -2\eta g + \eta^2 h g + \lambda \eta c \big[1 - \cos(\eta g)\big].$$

Indeed, the first centered iterate is $-\eta g$, and substitution into $f'(w) = g + hx + \lambda c(\cos x - 1)$ gives this expression. Writing $\Phi_\eta(g) = 1 - \cos(\eta g)$, each partition thus satisfies

$$D_2 = \eta^2 A + \lambda \eta \left[\frac{1}{K} \sum_i c_i \Phi_\eta(g_i) - \Phi_\eta(0.3) \right].$$

For the common control, the bracket is the same for rows and columns because their gradient multisets match. Hence $D_2^{\mathrm{row}} - D_2^{\mathrm{col}} = \eta^2 (A_{\mathrm{row}} - A_{\mathrm{col}})$, although either individual gap may contain a nonzero nonlinear correction. For the varying control, the exact between-partition correction is

$$D_2^{\mathrm{row}} - D_2^{\mathrm{col}} - \eta^2 (A_{\mathrm{row}} - A_{\mathrm{col}})$$
$$= \frac{\lambda \eta}{2MK} \sum_i (s_i - t_i) \Phi_\eta(0.3 + 2s_i). \tag{41}$$

Only the first two summands remain. With $g_i = 0.3 + 2s_i$, their expansion is

$$\frac{\lambda \eta^3}{4MK} (s_1 - s_2)(g_1^2 - g_2^2) + O(\eta^5).$$

Its leading coefficient is positive for the varying control, since $s_1 > s_2$, $g_1 - g_2 = 2(s_1 - s_2) > 0$, and $g_1 + g_2 = 0.6 + 2(s_1 + s_2) > 0$. Thus the nonquadratic correction is not identically zero, despite exact agreement of the input marginals.

The matching is exact at the common input. The FEMNIST analysis instead tests prediction on real-data objectives whose candidate marginals can differ; it does not rely on this construction's closed form.

Numerical check. The fixed grid uses $K \in \{3, 4, 8\}$, $\alpha \in \{.05, .1, .2, .4, .8\}$, $E \in \{1, 2, 5\}$, and all three controls, with $\eta = \alpha/(\max_{ij} h_{ij} + .75)$. All 135 deterministic cases satisfy the input-matching and one-step checks; the $E = 2$ endpoints also agree with the independent closed form. These are constructed cases, not statistical replications. Table 4 shows the designated $K = 4$, $\alpha = .8$, $E = 2$ comparison, where $\eta = 0.12973154$. The varying perturbation produces a resolved nonlinear difference, while the common perturbation cancels only between partitions.

Table 4: Nonquadratic controls on the same row/column construction. The last column is $D_2^{\mathrm{row}} - D_2^{\mathrm{col}} - \eta^2(A_{\mathrm{row}} - A_{\mathrm{col}})$. Values shown as zero have magnitude below 3×10^{-17}.

Perturbation	D_2^{row}	D_2^{col}	Nonlinear difference
Quadratic	0.02356238	0	0
Common	0.02571559	0.00215321	0
Varying	0.02601301	0.00195579	0.00049483

C A CANCELLATION-SENSITIVE TWO-STEP REMAINDER

The general bound in Appendix A.5 adds child and union Taylor errors. Bounding their difference directly can be tighter. The following elementary refinement focuses on the two-step map used in Theorem 1.

Proposition 4 (Two-step cancellation bound)**.** *Let $d = \sqrt{V_g(w)}$ and $\delta g_k = g_k(w) - \bar{g}(w)$. On a common convex region containing the input and first child and union iterates, suppose H_k is ρ-Lipschitz and $\Delta H_k = H_k - \bar{H}$ is ν_k-Lipschitz. Set $\nu = (\sum_k p_k \nu_k^2)^{1/2}$ and let $G \geq \|\bar{g}(w)\|$. Then, for $\eta \geq 0$,*

$$\|D_2(w) - \eta^2 A_{\mathcal{P}}(w)\| \leq \eta^3 \left(G\nu d + \frac{\rho}{2} d^2 \right). \tag{42}$$

Proof. Write $x = w - \eta \bar{g}(w)$ and $y_k = x - \eta \delta g_k$. Taylor expansion of each gradient about the *same* point x gives

$$\nabla F_k(y_k) = \nabla F_k(x) - \eta H_k(x) \delta g_k + a_k, \qquad \|a_k\| \leq \tfrac{1}{2} \rho \eta^2 \|\delta g_k\|^2.$$

Since $\sum_k p_k \nabla F_k(x) = \nabla \bar{F}(x)$ and $\sum_k p_k \delta g_k = 0$,

$$D_2 = \eta^2 \sum_k p_k \Delta H_k(x) \delta g_k - \eta \sum_k p_k a_k.$$

Subtract $\eta^2 A_{\mathcal{P}}(w)$ and use weighted Cauchy–Schwarz: the first difference has norm at most $\eta^2 \nu \|x - w\| d \leq \eta^3 G \nu d$; the remaining sum is bounded by $\eta^3 \rho d^2/2$. □

This bound is zero when all initial gradients agree, regardless of their common size. For quadratic children $\rho = \nu = 0$, recovering the exact two-step expression. The nonlinear term is also necessary: for equal-weight $F_\pm(x) = \lambda x^2/2 + \tau x^3/6 \pm \epsilon x$ at $w = 0$ on a bounded smooth region, $A = 0$ but $D_2 = -\tau \eta^3 \epsilon^2/2$, attaining the second term in Eq. 42.

Existing child-wise constants provide the safe choices $\rho = \max_k \rho_k$ and $\nu_k = \rho_k + \bar{\rho}$. For $E = 2$, taking the smaller of this remainder coefficient and $R_{\mathcal{P},2}$ preserves Eq. 8. The previously reported B^A and B^C tables use the original remainder; they are not silently recomputed with this refinement. For longer horizons, cancellation-sensitive bounds require control of derivative contrasts beyond the input. This two-step result does not replace the general-E bound.

D QUADRATIC MOMENTS AND FINITE LOCAL HORIZONS

The quadratic local-update polynomial is known through the surrogate-map formulation of Charles & Konečný (2021). We record its coefficient characterization, then construct partitions of the same atomic losses that match successively more coefficients. The construction also shows that the entire gradient-variance and two-step alignment fields can agree while later local-update maps differ.

Proposition 5 (Moments at a fixed input)**.** *Let $F_k(x) = x^\top H_k x/2 + b_k^\top x + c_k$, with $H_k = H_k^\top$, and let $p_k > 0$ sum to one. Fix w and define $g_k = H_k w + b_k$, $\bar{H} = \sum_k p_k H_k$, $\bar{g} = \sum_k p_k g_k$, and $M_r(\mathcal{P}; w) = \sum_k p_k H_k^r g_k$. Let $P_E^{\mathcal{P}}(w; \eta)$ average the child endpoints after E full-gradient steps of constant step size η, all starting at w. Let $U_E(w; \eta)$ be the corresponding endpoint on $\bar{F} = \sum_k p_k F_k$ and $D_E = P_E^{\mathcal{P}} - U_E$. For every integer $E \geq 1$ and real η,*

$$P_E^{\mathcal{P}}(w; \eta) = w + \sum_{j=1}^{E} \binom{E}{j} (-\eta)^j M_{j-1}(\mathcal{P}; w), \tag{43}$$

$$D_E(\mathcal{P}; w, \eta) = \sum_{j=2}^{E} \binom{E}{j} (-\eta)^j \big[M_{j-1}(\mathcal{P}; w) - \bar{H}^{j-1}\bar{g}\big]. \tag{44}$$

Two decompositions of the same aggregate objective have identical E-step endpoints at this fixed input throughout a nonempty open interval of η if and only if their M_r agree for $1 \leq r < E$.

Proof. The gradient after t steps on child k is $(I - \eta H_k)^t g_k$, giving displacement $-\eta \sum_{t=0}^{E-1} (I - \eta H_k)^t g_k$. Expand the matrix powers and use $\sum_{t=r}^{E-1} \binom{t}{r} = \binom{E}{r+1}$ to obtain (43). Apply it to $\bar{F}$ and cancel $j = 1$ to obtain (44). For two decompositions of the same objective, the endpoint difference is the vector polynomial $\sum_{j=2}^{E} \binom{E}{j} (-\eta)^j [M_{j-1}(\mathcal{P}; w) - M_{j-1}(\mathcal{Q}; w)]$. It vanishes on an open interval exactly when all coefficients vanish. These finite identities require neither invertible Hessians nor commutation between different children, and $E = 1$ is the empty condition. □

For $E \geq 3$, equality at one step size need not identify the coefficients. At input zero, take equal weights, curvatures $(1, 3/2)$, and gradients $(1, -1)$; reverse the gradients for the second decomposition. Their first two moments are $(-1/4, -5/8)$ and $(1/4, 5/8)$, respectively, yet both three-step defects vanish at $\eta = 6/5$, since the first defect is $-3\eta^2/4 + 5\eta^3/8$ and the second is its negative.

Proposition 6 (Matching moments before an arbitrary horizon)**.** *For every integer $m \geq 1$ and $\beta > 0$, there are two equal-size partitions $\mathcal{P}_0, \mathcal{P}_1$ of the same strictly convex scalar quadratic atoms, with $K = 2^{m+1}$ children each, such that their complete separate gradient and Hessian multisets agree at input zero and $M_r(\mathcal{P}_0; 0) = M_r(\mathcal{P}_1; 0) = 0$ for $0 \leq r < m$. Their common aggregate gradient is zero there. For every real η, both defects vanish through horizon m; $D_e(\mathcal{P}_0; 0, \eta) = 0$ for every $e \geq 1$, whereas*

$$D_{m+1}(\mathcal{P}_1; 0, \eta) = -\eta^{m+1} \beta^m m!\, 2^{m(m-1)/2 - m}. \tag{45}$$

Proof. Put $n = 2^m$ and choose $h_0 > \beta(n-1)/2$. Index the $K = 2n$ children by $i = (j, a)$, with $0 \leq j < n$ and $a \in \{0, 1\}$. Define

$$h_i = h_0 + \beta j, \qquad g_i^0 = (-1)^a, \qquad g_i^1 = s_j := (-1)^{\mathrm{popcount}(j)}.$$

Give each child weight $1/K$ and loss $F_i^b(x) = h_i x^2/2 + g_i^b x$. The Hessian lists coincide. Since $s_{2j+1} = -s_{2j}$, each gradient list contains n entries of each sign; both mean gradients are zero, and both mean Hessians are $\bar{h} = h_0 + \beta(n-1)/2$. Pairwise cancellation gives $M_r(\mathcal{P}_0; 0) = 0$ for all r. For $\mathcal{P}_1$, the moment is $n^{-1} \sum_{j=0}^{n-1} s_j (h_0 + \beta j)^r$. Binary expansion gives the classical finite-product identity

$$\sum_{j=0}^{n-1} s_j e^{(h_0 + \beta j)t} = e^{h_0 t} \prod_{q=0}^{m-1} (1 - e^{\beta 2^q t}). \tag{46}$$

The product has a zero of order m at $t = 0$ and leading coefficient $(-1)^m \beta^m 2^{m(m-1)/2}$. Taking derivatives proves vanishing for $r < m$ and $M_m(\mathcal{P}_1; 0) = (-1)^m \beta^m m! 2^{m(m-1)/2 - m}$. The union stays at zero, so (43) gives the horizon claims. At $E = m + 1$, only the last term survives, with sign $(-1)^{m+1}(-1)^m = -1$.

For a shared-data realization, use K^2 equal-mass labeled atoms

$$\ell_{i\ell}(x) = \tfrac{1}{2}(h_i + h_\ell - \bar{h})x^2 + (g_i^0 + g_\ell^1)x.$$

Their row means are F_i^0 and column means are F_ℓ^1. Every atom has curvature at least $h_0 - \beta(n - 1)/2 > 0$. Both partitions have K atoms per child and exactly the same union objective. □

Identical variance and alignment fields. For this construction, put $V_H = K^{-1}\sum_i (h_i - \bar{h})^2 = \beta^2(n^2-1)/12$ and $\mu_1^b = K^{-1}\sum_i h_i g_i^b$. At any input w, the child gradient is $h_i w + g_i^b$, so

$$V_g^b(w) := K^{-1}\sum_i [(h_i - \bar{h})w + g_i^b]^2 = V_H w^2 + 2w\mu_1^b + 1,$$
$$A^b(w) := K^{-1}\sum_i (h_i - \bar{h})[(h_i - \bar{h})w + g_i^b] = V_H w + \mu_1^b = \tfrac{1}{2}(V_g^b)'(w). \tag{47}$$

For $m \geq 2$, both μ_1^b vanish: the entire V_g and A fields coincide. Moreover, identical Hessian multisets make the linear coefficients of the two affine endpoint maps equal at every horizon. Their maps therefore agree for every input through horizon m, and their difference at $m+1$ is the input-independent constant in (45). This does not assert equal gradient multisets away from zero.

For example, $m = 2$, $h_0 = 2$, and $\beta = 1$ give

$$(h_i) = (2, 2, 3, 3, 4, 4, 5, 5),$$
$$(g_i^0) = (1, -1, 1, -1, 1, -1, 1, -1),$$
$$(g_i^1) = (1, 1, -1, -1, -1, -1, 1, 1).$$

These two partitions of 64 strictly convex atoms have $V_g(w) = 5w^2/4 + 1$ and $A(w) = 5w/4$ for every w, but $P_3^{\mathcal{P}_1}(w;\eta) - P_3^{\mathcal{P}_0}(w;\eta) = -\eta^3$.

Taking $m = E - 1$ establishes separation at any prescribed $E \geq 2$ after matching all earlier cross-moments. Adding $\lambda \sum_{a=2}^d x_a^2/2$, $\lambda > 0$, to each atom embeds the scalar discrepancy along the unit vector e_1; its Euclidean norm is $|\eta|^{m+1}\beta^m m! 2^{m(m-1)/2-m}$. This existence construction uses exponentially many children and does not give a uniform-in-horizon gap or a practical estimator. It can use stable steps: $h_0 = 1$, $\beta = 1/(2^m - 1)$, and $\eta = 1/4$ keep child curvatures in $[1, 2]$ and atomic curvatures in $[1/2, 5/2]$.

E A FINITE-SUPPORT BOUNDARY AND SHARP DELAYED DISCREPANCY

The quadratic update polynomial is known (Charles & Konečný, 2021). Here classical Vandermonde interpolation gives a finite identification boundary, and classical Chebyshev approximation gives a sharp discrepancy at that boundary. Translate the common reference input to zero. Two scalar quadratic decompositions share a weighted curvature distribution $\sum_{j=1}^q p_j \delta_{h_j}$, with distinct $h_j > 0$, $p_j > 0$, and $\sum_j p_j = 1$. Write g_j^b for their conditional mean input gradients at curvature h_j, where $b \in \{0, 1\}$, and set $M_r^b = \sum_j p_j h_j^r g_j^b$. Let $P_E^b(w;\eta)$ average the endpoints after E full-gradient steps of constant step size η from input w. Summing each child's gradients $(1 - \eta h_j)^t g_j^b$ gives

$$P_E^b(0;\eta) = \sum_{\ell=1}^{E} \binom{E}{\ell} (-\eta)^\ell M_{\ell-1}^b. \tag{48}$$

Coalescing children of equal curvature through their conditional mean gradients preserves these maps.

Proposition 7 (Identification on a fixed curvature support)**.** *For any fixed $\eta_0 \neq 0$, the following are equivalent: the moments $M_0, \ldots, M_{q-1}$ match; the endpoints $P_E^0(0;\eta_0)$ and $P_E^1(0;\eta_0)$ match for $E = 1, \ldots, q$; the conditional means g_j^0, g_j^1 match at every curvature; and the two averaged maps agree at every input, step size, and finite horizon.*

Proof. The q moment differences are the Vandermonde matrix $(h_j^r)_{0 \leq r < q,\, 1 \leq j \leq q}$ times $(p_j(g_j^1 - g_j^0))_j$. Distinct nodes make this matrix invertible. Equation (48) at horizons $1, \ldots, q$ is triangular in these moments, with nonzero diagonal $(-\eta_0)^E$. These observations prove the first three equivalences. Equal conditional means give all moments and all endpoints equal at zero. At input w, each conditional gradient becomes $g_j^b + h_j w$; the common weighted curvature distribution makes the additional moments identical, proving the all-input statement. The reverse implication is immediate. □

If the aggregate objective is already fixed, M_0 and horizon one are redundant. For $q = 1$ the proposition reduces to equality of the average gradient. The following separation result assumes $q \geq 2$.

Proposition 8 (Sharp discrepancy after matching lower moments)**.** *Suppose $h_j \in [a, b]$, $0 < a < b$, $\Delta = b - a$, and all child input gradients have magnitude at most $G > 0$. If $M_r^0 = M_r^1$ for $0 \leq r \leq q - 2$, then, for every input w,*

$$|P_q^1(w; \eta) - P_q^0(w; \eta)| \leq B_q := \frac{G|\eta|^q \Delta^{q-1}}{2^{2q-4}}. \tag{49}$$

This is sharp over the common support locations and weights in $[a, b]$, even if the full weighted input-gradient distributions must match. An equality case has q children per decomposition and, for every $\eta \neq 0$, first differs at horizon q. It is realizable by two partitions of the same strictly convex quadratic atoms for every positive interval $0 < a < b$.

Proof. Put $n = q - 1$ and $C_n = \Delta^n / 2^{2n-1}$. The signed measure $\nu = \sum_j p_j (g_j^1 - g_j^0) \delta_{h_j}$ has total variation at most $2G$ and annihilates every polynomial of degree less than n. The classical monic Chebyshev polynomial on this interval is

$$P_n(t) = C_n T_n\left(\frac{2t - a - b}{\Delta}\right), \qquad \|P_n\|_{\infty,[a,b]} = C_n.$$

Indeed, T_n has leading coefficient 2^{n-1} and satisfies $T_n(\cos\theta) = \cos(n\theta)$. The monic minimax property is classical (NIST, 2026a); only this explicit polynomial is needed for the bound:

$$|M_n^1 - M_n^0| = \left| \int P_n \, d\nu \right| \leq 2GC_n.$$

All earlier terms cancel in (48), yielding (49). Since the curvature distribution is common, the difference of the two affine maps is input-independent.

To attain equality, take the Chebyshev–Lobatto nodes and weights

$$\begin{aligned} h_j &= \frac{a+b}{2} + \frac{\Delta}{2}\cos(j\pi/n), \qquad 0 \leq j \leq n, \\ p_0 &= p_n = \frac{1}{2n}, \qquad p_j = \frac{1}{n} \ (0 < j < n), \\ g_j^0 &= G(-1)^j, \qquad g_j^1 = -G(-1)^j. \end{aligned} \tag{50}$$

Here is a direct verification of the classical signed-weight cancellation. For $\theta_j = j\pi/n$, put $S_r = \sum_j p_j \cos(r\theta_j)$. Reflection gives $S_r = 0$ for odd r; for even $r = 2m$ with $0 < m < n$, the endpoint-halved sum equals $n^{-1} \operatorname{Re} \sum_{j=0}^{n-1} \exp(2\pi i m j/n) = 0$. Also $S_0 = S_{2n} = 1$. Using $(-1)^j = \cos(n\theta_j)$ and the product-to-sum formula therefore gives

$$\sum_j p_j (-1)^j T_r(\cos\theta_j) = 0 \quad (r < n), \qquad \sum_j p_j (-1)^j T_n(\cos\theta_j) = 1.$$

Thus $M_r^0 = M_r^1 = 0$ for $r < n$, $M_n^0 = GC_n$, and $M_n^1 = -GC_n$. In particular, each weighted gradient law assigns mass $1/2$ to each of $\{-G, G\}$. The first signed map difference is

$$P_q^1(w; \eta) - P_q^0(w; \eta) = (-1)^{q+1} \frac{G\eta^q \Delta^{q-1}}{2^{2q-4}}. \tag{51}$$

Finally, the mean curvature is $c = (a + b)/2$, and the mean gradients are zero. Give atom (i, j) mass $p_i p_j$ and loss

$$\ell_{ij}(x) = \frac{h_i h_j}{2c} x^2 + (g_i^0 + g_j^1)x.$$

Since $\sum_j p_j h_j = c$ and both mean gradients vanish, the normalized row means are $h_i x^2/2 + g_i^0 x$, and column means are $h_j x^2/2 + g_j^1 x$. Both unions equal $cx^2/2$. Every atomic curvature lies in $[a^2/c, b^2/c]$, whose lower endpoint is positive whenever $a > 0$. □

Prescribed support and weights are a smaller class. For fixed h_j, p_j, let $\omega(t) = \prod_j (t - h_j)$. Lagrange interpolation gives $\sum_j h_j^r/\omega'(h_j) = 0$ for $r < q-1$ and 1 for $r = q-1$. The lower-moment nullspace is consequently $p_j(g_j^1 - g_j^0) = \lambda/\omega'(h_j)$, with $\Delta M_{q-1} = \lambda$. The exact optimum in this class is

$$\max |\Delta M_{q-1}| = 2G \min_j p_j |\omega'(h_j)|,$$

attained by $g_j^1 = -g_j^0 = \lambda/(2p_j\omega'(h_j))$. This optimum does not impose matching full gradient laws and can be strictly smaller than the interval-wide Chebyshev bound.

Corollary 3 (Prediction from the fixed Lobatto observations)**.** *Fix the weighted curvature law in* (50)*, the input-gradient law assigning mass* $1/2$ *to each of* $\{-G, G\}$*, and the observed moments* $M_0 = \cdots = M_{q-2} = 0$*. Over all compatible joint laws, the exact minimax absolute error for predicting* $P_q(0;\eta)$ *is* $B_q/2$*, attained by prediction zero. At input* w*, a minimax predictor is* $w\sum_j p_j(1-\eta h_j)^q$*, with the same error.*

Proof. For any compatible joint law, its conditional gradients have magnitude at most G. Monic Chebyshev subtraction, now for one law rather than a difference of two laws, gives $|M_{q-1}| \le G\Delta^{q-1}/2^{2q-3}$. Equation (48) bounds the endpoint by $B_q/2$. The two sign assignments in (50) are compatible and attain opposite endpoints of this interval. Any common prediction has maximum error at least half their separation, so the zero prediction is minimax. The common curvature law fixes the affine map's slope at other inputs. □

Global fields and representation weights. For $q \ge 3$, the construction has $V_H = \Delta^2/8$ and $\sum_j p_j h_j g_j^b = 0$. Expanding the squared gradient deviation gives, for every input,

$$V_g^b(w) = V_H w^2 + G^2, \qquad A^b(w) = V_H w.$$

Thus the entire variance and alignment fields agree despite (51). For example, $h = (2, 3/2, 1)$, $p = (1/4, 1/2, 1/4)$, and gradients $(G, -G, G)$ versus their negatives give $V_g(w) = G^2 + w^2/8$, $A(w) = w/8$, and $P_3^1 - P_3^0 = G\eta^3/4$. The weighted gradient laws match, but these unweighted three-child gradient lists do not. At $q = 2$, the first moments are $\pm G\Delta/2$ and the global fields instead differ; the gap bound still holds.

Writing $p_j = d_j/(2n)$, with endpoint $d_j = 1$ and interior $d_j = 2$, replication of atom (i, j) by $d_i d_j$ copies yields $4n^2$ equal-mass atoms while retaining q unequal-size children. Alternatively, splitting interior children into two identical copies yields $2n$ equal-weight children and the same equal-mass grid; their ordinary gradient lists then also match. No minimal-child-count claim is made. The gradient bound G is on children at the reference input, not atoms or entire trajectories. Finally, $B_q = 4G|\eta|(|\eta|\Delta/4)^{q-1}$ is geometrically small under $0 < \eta b \le 1$. For $[a, b] = [1, 2]$ and $\eta = 1/4$, the equality-case gap is $G/16^{q-1}$, with atomic curvatures in $[2/3, 8/3]$.

F Finite-spectrum identification without commuting Hessians

This section proves Theorem 2, its sharpness under same-atom constraints, and the vector version of Eq. (59). Spectral decomposition, Vandermonde inversion, and Chebyshev approximation are classical; the constructions specify the information shared by the competing partitions.

F.1 Endpoint observations and spectral weights

Translate the reference input to zero. Write $F_k^b(x) = \frac{1}{2}x^\top H_k^b x + (g_k^b)^\top x + c_k^b$, with positive weights summing to one and symmetric positive-definite Hessians. Let Λ be the union of their spectra, $q = |\Lambda|$, and let $\Pi_{k,\lambda}^b$ be the orthogonal spectral projectors, zero when an eigenvalue is absent. Define

$$W_\lambda^b = \sum_k p_k^b \Pi_{k,\lambda}^b, \qquad R_\lambda^b = \sum_k p_k^b \Pi_{k,\lambda}^b g_k^b.$$

These weights satisfy $W_\lambda^b \succeq 0$, $\sum_\lambda W_\lambda^b = I$, and $\sum_\lambda \lambda W_\lambda^b = \bar{H}^b$. In general they are weighted sums of projectors, not projectors themselves. For $\phi_E(h;\eta) = -\eta\sum_{t=0}^{E-1}(1-\eta h)^t$, the exact affine

map is

$$P_E^b(w;\eta) = S_E^b(\eta)w + y_E^b(\eta),$$
$$S_E^b(\eta) = \sum_\lambda (1-\eta\lambda)^E W_\lambda^b, \qquad y_E^b(\eta) = \sum_\lambda \phi_E(\lambda;\eta) R_\lambda^b. \tag{52}$$

It follows directly by applying the known quadratic map to each child (Appendix I.1). No products between different children's Hessians are involved.

At horizons $1, \dots, q$, the endpoint polynomial is triangular in $M_r^b = \sum_\lambda \lambda^r R_\lambda^b$, $0 \le r < q$, with diagonal coefficients $(-\eta_0)^E$. For $\eta_0 \neq 0$, equal endpoints therefore give equal moments. The scalar Vandermonde matrix $(\lambda^r)_{0\le r<q, \lambda\in\Lambda}$ is invertible and acts coordinatewise, so the vector residues coincide. Conversely, equal residues give equal endpoints at zero for every horizon and step size. If the spectral weights also coincide, Eq. (52) has the same slope in both decompositions and proves all-input equality. This establishes the identification part of the theorem.

The weights are an exact description of the missing slopes: equality of $S_E^0(\eta_0)$ and $S_E^1(\eta_0)$ for $E = 0, \dots, q-1$ is equivalent, by the same two inversions, to equality of every W_λ. Here $S_0 = I$ is automatic. When the residues at zero already agree, the two maps agree for every horizon and step size precisely on the input subspace

$$\bigcap_{\lambda\in\Lambda} \ker(W_\lambda^1 - W_\lambda^0).$$

Indeed, at w the residue difference is $\lambda(W_\lambda^1 - W_\lambda^0)w$; positivity of λ and Vandermonde inversion give necessity as well as sufficiency.

Without common spectral weights, matching the first q horizons at $d+1$ affinely independent inputs identifies every later map: first apply the fixed-input result at each input, then use uniqueness of an affine map on an affine basis. This is a sufficient test, not a minimal-query claim. Also, q can be as large as $d(K_0 + K_1)$. The logical identification result concerns the actual union spectrum; constructing an extrapolator additionally requires the support or equivalent recurrence information.

F.2 Dimension-free bound and noncommuting equality cases

Put $n = q-1$ and $c = (a+b)/2$. The monic polynomial

$$T_n^*(h) = \frac{\Delta^n}{2^{2n-1}} T_n\left(\frac{2h-a-b}{\Delta}\right)$$

has uniform norm $C_n = \Delta^n/2^{2n-1}$ on $[a,b]$ (NIST, 2026a). Matching the first $q-1$ horizons cancels all moments of degree below n, leaving

$$\Delta M_n = \sum_k p_k^1 T_n^*(H_k^1) g_k^1 - \sum_k p_k^0 T_n^*(H_k^0) g_k^0.$$

The spectral theorem gives $\|T_n^*(H_k^b)\|_{\mathrm{op}} \le C_n$. Applying the triangle inequality to children, each with $\|g_k^b\| \le G$, yields $\|\Delta M_n\| \le 2GC_n$. The first remaining endpoint term is $(-\eta)^q \Delta M_n$, proving Eq. (6) at zero. Common spectral weights make the map difference input-independent. This argument has no dimension factor; bounding the sum of norms of spectral vector residues instead would introduce an unnecessary loss.

For sharpness, use the nodes h_j, weights p_j, and signs $s_j = (-1)^j$ of Eq. (50). Their signed moments vanish below n and equal C_n at degree n. Choose two noncommuting rank-one orthogonal projectors in dimension two,

$$P_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad P_1 = \frac{1}{25}\begin{pmatrix} 9 & 12 \\ 12 & 16 \end{pmatrix}, \qquad Q_t = I - P_t,$$

and a unit vector v. Child (t,j,k) has mass $p_j p_k/2$ and

$$H_{tjk} = h_j P_t + h_k Q_t, \qquad g_{tjk}^0 = G(s_j P_t + s_k Q_t)v, \qquad g_{tjk}^1 = -g_{tjk}^0.$$

The sign matrix is orthogonal, so every gradient has norm G. Summing over the unused node gives $W_{h_j} = p_j I$ in both partitions, including children with repeated eigenvalues. Their mean Hessian is cI, and

$$M_r^0 = G\left(\sum_j p_j s_j h_j^r\right) v = -M_r^1.$$

Thus every earlier moment cancels and the signed first gap is $(-1)^{q+1} B_q v$ for $\eta > 0$ (replace $|\eta|^q$ by η^q for arbitrary signed steps). The Hessians $aI + \Delta P_0$ and $aI + \Delta P_1$ have nonzero commutator $\Delta^2[P_0, P_1]$; the example is not a simultaneously diagonal scalar embedding.

The complete vector-gradient laws also coincide. Under $p_j p_k$, each sign pair occurs with probability $1/4$, because the scalar signed mass is zero. For a fixed orientation, the gradient distribution is uniform on Gv, $G(2P_t - I)v$, $-G(2P_t - I)v$, and $-Gv$, counting coincident vectors with their masses. It is invariant under negation; mixing the two orientations preserves this equality.

Both partitions arise from the same strictly convex losses. Atom (t, j, k, l, m) has mass $p_j p_k p_l p_m/2$, Hessian and gradient

$$H^{\text{atom}}_{tjklm} = \frac{h_j h_l}{c} P_t + \frac{h_k h_m}{c} Q_t, \qquad g^{\text{atom}}_{tjklm} = g^0_{tjk} + g^1_{tlm}.$$

Grouping by (t, j, k) or by (t, l, m) gives the stated child means, using $\sum_j p_j h_j = c$ and the zero mean gradient within each orientation. Every atomic eigenvalue lies in $[a^2/c, b^2/c]$, so strict convexity holds for every $0 < a < b$. There are $2q^4$ weighted atoms; replication produces $32(q-1)^4$ equal-mass atoms. This is an existence construction, not a bound claimed sharp in the number of atoms.

On the fixed spectral and vector-gradient laws of this construction, with $M_0 = \cdots = M_{q-2} = 0$, the minimax Euclidean error for predicting $P_q(0; \eta)$ is exactly $B_q/2$. The single-decomposition argument above bounds its norm by $GC_n|\eta|^q = B_q/2$, so zero is an upper-bound predictor. The two antipodal equality cases prove the matching lower bound. At another input the common slope supplies the prediction center.

F.3 A COMPACT THREE-EIGENVALUE EXAMPLE

For $q = 3$, eight equal-weight children and 32 equal-mass atoms suffice. In each orientation t, use four children with

i	1	2	3	4
(u_i, v_i)	(a, b)	(b, a)	(c, c)	(c, c)
σ_i	1	1	-1	-1

and $H_{ti} = u_i P_t + v_i Q_t$, $g^0_{ti} = G\sigma_i v = -g^1_{ti}$. Each partition has four $+Gv$ and four $-Gv$ gradients, with $W_a = W_b = I/4$ and $W_c = I/2$. For each (t, i, j), give the atom mass $1/32$, Hessian $(u_i u_j/c)P_t + (v_i v_j/c)Q_t$, and gradient $G(\sigma_i - \sigma_j)v$. Rows and columns each contain four atoms. Direct averaging gives

$$M_0 = M_1 = 0, \quad M_2^0 = G\Delta^2 v/8 = -M_2^1, \qquad P_3^1 - P_3^0 = G\eta^3\Delta^2 v/4.$$

Both partitions also have identical entire fields

$$V_g(w) = G^2 + \frac{\Delta^2}{8}\|w\|^2, \qquad A_{\mathcal{P}}(w) = \frac{\Delta^2}{8} w.$$

These follow from the shared second Hessian moment and $M_0 = M_1 = 0$. For $[a, b] = [1, 2]$, $G = 1$, $v = e_1$, the first gap is $e_1/256$ at $\eta = 1/4$, and atomic eigenvalues lie in $[2/3, 8/3]$.

F.4 SAME SPECTRAL HISTOGRAM, DIFFERENT INPUT DEPENDENCE

Use five children of mass $1/5$ in each decomposition. The common child has $H_c = \left(\begin{smallmatrix} 4 & 1 \\ 1 & 4 \end{smallmatrix}\right)$ and $g_c = e_1$. The other four children are a $+e_2, -e_2$ gradient pair at each of two Hessians:

$$\mathcal{P}_0 : \ \text{diag}(3, 4), \text{diag}(5, 4), \qquad \mathcal{P}_1 : \ \text{diag}(4, 3), \text{diag}(4, 5).$$

The union quantities are $\bar{H} = \left(\begin{smallmatrix} 4 & 1/5 \\ 1/5 & 4 \end{smallmatrix}\right)$, $\bar{g} = e_1/5$. The full gradient lists and weighted eigenvalue histograms are identical; both decompositions contain noncommuting Hessians. At zero, each signed gradient pair cancels at every horizon, leaving in both decompositions

$$P_E(0;\eta) = \tfrac{1}{5} H_c^{-1}[(I - \eta H_c)^E - I]e_1.$$

It is nonzero for $E \geq 1$ and $0 < \eta < 1/5$. The common child's projectors cancel between the decompositions. With $T = (2/5)\,\mathrm{diag}(-1, 1)$, the remaining children give

$$\Delta W_3 = \tfrac{2}{5}(e_2 e_2^\top - e_1 e_1^\top) = T,$$
$$\Delta W_4 = \tfrac{4}{5}(e_1 e_1^\top - e_2 e_2^\top) = -2T,$$
$$\Delta W_5 = \tfrac{2}{5}(e_2 e_2^\top - e_1 e_1^\top) = T.$$

Their traces and first two matrix moments (degrees zero and one) vanish, but the degree-two moment is $2T$. Consequently

$$P_2^1(w;\eta) - P_2^0(w;\eta) = \tfrac{4}{5}\eta^2\,\mathrm{diag}(-1, 1)w.$$

To realize these children on the same data, assign mass $1/25$ to each row-column atom with Hessian $H_i^0 + H_j^1 - \bar{H}$ and gradient $g_i^0 + g_j^1 - \bar{g}$. Row and column averages recover the specified losses. Since each child Hessian is at least $3I$ and $\bar{H} \preceq (21/5)I$, every atom is at least $(9/5)I$.

F.5 Perturbed vector observations

Fix $\eta \neq 0$ and distinct positive nominal nodes $\widehat{h}_j$. As in Appendix H, divide z^L by $(z-1)\prod_j[z - (1 - \eta\widehat{h}_j)]$ to obtain coefficients $a_{L,m}$. The polynomial $J_L(h) = \sum_{m=1}^q a_{L,m}\phi_m(h;\eta)$ interpolates ϕ_L at the nominal nodes. Its scalar residual satisfies

$$|r_L(h)| := |\phi_L(h;\eta) - J_L(h)| \leq K_{L,q}(\rho)\prod_j |h - \widehat{h}_j|$$

on the compact positive interval containing actual and nominal spectra (NIST, 2026b, Section 3.3(i)). The vector endpoint prediction error before observation noise is exactly $\sum_k p_k r_L(H_k) g_k(w_0)$. For each symmetric Hessian, $\|r_L(H_k)\|_{\mathrm{op}} = \max_{\lambda \in \mathrm{spec}(H_k)} |r_L(\lambda)|$. Applying the child-level norm bound and then the triangle inequality for observation errors proves Eq. (59) without a dimension factor or commutativity assumption. The actual spectrum need not have q values.

For independent Euclidean error balls of radii ε_m, the worst-case contribution of observation noise to this fixed linear predictor is exactly $\sum_m |a_{L,m}|\varepsilon_m$: align all errors with one unit vector and the coefficient signs. This concerns the specified predictor, not an optimal estimator under extra feasibility constraints. As an exact check, a rotated Hessian with eigenvalues $5/4, 7/4$, nominal nodes $(1, 3/2, 2)$, $L = 4$, and $\eta = 1/4$ gives residual $(H - I)(H - \frac{3}{2}I)(H - 2I)e_1/256$, attaining norm $3/16384$.

G Sharp approximate identification from noisy horizons

Exact spectrum counts do not determine how many noisy observations are useful. This section gives the uniform endpoint diameter for a declared spectral uncertainty set and observation-error budget. The optimization and its measure dual are instances of classical optimal recovery (Ettehad & Foucart, 2020, Theorem 4); the additional construction realizes its extremizers as two partitions of one strictly convex weighted loss pool.

G.1 Information and sharp radius

Fix a nonempty compact set $S \subset (0, \infty)$, a child-gradient bound $G > 0$, a step size $\eta \neq 0$, a target horizon L, and $1 \leq r < L$. The two decompositions have the same aggregate quadratic objective. At the reference input w_0, their child gradients have norm at most G; every child-Hessian eigenvalue belongs to S. Write

$$\phi_m(h) = -\eta \sum_{t=0}^{m-1} (1 - \eta h)^t.$$

For $m = 2, \dots, r$, the same recorded endpoint can differ from either true endpoint by at most ε_m in Euclidean norm. Define

$$\mathcal{E}_r = \inf_{b\in\mathbb{R},\ a\in\mathbb{R}^{r-1}} \left\{ G \left\| \phi_L - b - \sum_{m=2}^{r} a_m \phi_m \right\|_{\infty,S} + \sum_{m=2}^{r} |a_m| \varepsilon_m \right\}. \tag{53}$$

The constant is free because the aggregate gradients agree. The first endpoint therefore adds no constraint on a difference between these decompositions. All norms and error radii refer to this fixed step size.

Theorem 3 (Sharp noisy-horizon diameter)**.** *Consider quadratic decompositions with a common objective, child spectra in S, and child gradients of norm at most G at w_0. If their endpoints at $m = 2, \dots, r$ admit a common recorded transcript with error at most ε_m on each side, then*

$$\|P_L^1(w_0) - P_L^0(w_0)\| \le 2\mathcal{E}_r.$$

The bound is sharp over finite weighted loss pools. Whenever $\mathcal{E}_r > 0$, equality is attained by two partitions of one strictly convex two-dimensional pool with noncommuting child Hessians, common matrix spectral weights, and the same reference-point gradient law. The attaining pool and its common objective are chosen by the construction. Thus the smallest depth uniformly forcing every compatible pair to be τ-close is $\min\{r < L : 2\mathcal{E}_r \le \tau\}$, if it exists.

The largest possible horizon-L separation is $2\mathcal{E}_r$. This statement is uniform over finite weighted quadratic loss pools satisfying the declared bounds. It does not assert that every supplied dataset, objective, or child-Hessian list attains this diameter. When $\mathcal{E}_r > 0$, equality is realized in dimension two by noncommuting children, with both partitions sharing their complete child-Hessian list, masses, matrix-valued spectral weights, and reference-point gradient distribution. The realizing atomic weights can be positive real numbers; an equal-example-mass restriction is not imposed here.

Upper bound. Put $\Delta y_m = (P_m^1(w_0) - P_m^0(w_0))/2$. For any coefficients in Eq. (53), set $R = \phi_L - b - \sum_{m=2}^{r} a_m \phi_m$. Equal aggregate gradients imply

$$\Delta y_L - \sum_{m=2}^{r} a_m \Delta y_m = \frac{1}{2} \left[\sum_k p_k^1 R(H_k^1) g_k^1 - \sum_k p_k^0 R(H_k^0) g_k^0 \right].$$

The spectral theorem bounds the right side by $G\|R\|_{\infty,S}$, without a common eigenbasis. The common transcript gives $\|\Delta y_m\| \le \varepsilon_m$. Taking the infimum proves the bound. If the matrix-valued spectral weights also agree, the two maps have the same linear part, so their difference is constant across all inputs.

Classical dual problem. For finite signed measures on S, convex duality gives

$$\mathcal{E}_r = \max_{\nu} \left\{ \int \phi_L \, d\nu : \|\nu\|_{\mathrm{TV}} \le G, \quad \nu(S) = 0, \quad \left| \int \phi_m \, d\nu \right| \le \varepsilon_m \ (2 \le m \le r) \right\}. \tag{54}$$

For finite S, this is linear-program duality after splitting the positive and negative parts of ν. For compact S, separation of the compact moment body gives the same identity. One may choose a finite-support maximizer: represent ν/G as a signed marginal of a probability measure on $S \times \{-1, 1\}$, filling unused total variation with cancelling mass, and preserve its mass, observed moments, and target moment by finite-dimensional convex-hull reduction. The feasible set is symmetric, so maximizing the absolute target gives the same value.

For the zero-mass signed-measure model, Eq. (53) is also the minimax recovery error, not just the best error of a prescribed interpolant. The upper bound is linear recovery. A maximizing ν and $-\nu$ can both produce the zero transcript with admissible, opposite errors, while their target values are $\pm\mathcal{E}_r$. Every predictor errs by at least $\mathcal{E}_r$ on one of them. The following construction preserves this ambiguity under the paper's same-pool constraint.

G.2 Realizing the extremizer with noncommuting children

Let $t = \|\nu\|_{\mathrm{TV}}/G \le 1$ for a finite maximizer and choose $h_0 \in S$. The positive measures

$$\alpha_+ = \nu_+/G + (1-t)\delta_{h_0}/2, \qquad \alpha_- = \nu_-/G + (1-t)\delta_{h_0}/2$$

each have mass $1/2$. Take a finite probability coupling γ with marginals $2\alpha_+$ and $2\alpha_-$. Its pairs (h_+^e, h_-^e) and masses γ_e satisfy, for every scalar function f,

$$\frac{G}{2}\sum_e \gamma_e[f(h_+^e) - f(h_-^e)] = \int f\, d\nu. \tag{55}$$

For each pair introduce four children, of mass $\gamma_e/4$, at angles $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$. With Q_θ the planar rotation, let

$$H_{e\theta} = Q_\theta \operatorname{diag}(h_+^e, h_-^e) Q_\theta^\top, \quad S_\theta = Q_\theta \operatorname{diag}(1, -1) Q_\theta^\top, \quad g_{e\theta}^{\pm} = \pm G S_\theta e_1.$$

The vectors $S_\theta e_1$ are $e_1, e_2, -e_1, -e_2$. Both gradient distributions are consequently uniform on these four vectors multiplied by G, and have zero mean. The Hessians and masses are identical on the two sides. Direct multiplication yields

$$f(H_{e\theta}) G S_\theta e_1 = \frac{G}{2}\{[f(h_+^e) - f(h_-^e)]I + [f(h_+^e) + f(h_-^e)]S_\theta\} e_1.$$

Since $\sum_\theta S_\theta = 0$, Eq. (55) gives

$$\sum_{e,\theta} \frac{\gamma_e}{4} f(H_{e\theta}) g_{e\theta}^{\pm} = \pm \left(\int f\, d\nu \right) e_1.$$

Their common aggregate Hessian is cI, where $c = \int h\, d(\alpha_+ + \alpha_-) > 0$. Their matrix spectral weights are $W_h = (\alpha_+ + \alpha_-)(\{h\})I$. If $\nu \neq 0$, a pair with $h_+ \neq h_-$ exists; its Hessians at angles zero and $\pi/4$ do not commute. Setting $f = \phi_m$ produces the required early offsets and the target separation $2\mathcal{E}_r$.

One loss pool for both partitions. For any finite SPD list H_i with masses $p_i > 0$, $\sum_i p_i = 1$, and $\sum_i p_i g_i = 0$, choose $0 < \zeta < \min_i \lambda_{\min}(H_i)$. Use atom (i, j) of mass $p_i p_j$, Hessian A_{ij}, and input gradient b_{ij}, where

$$A_{ij} = \begin{cases} \zeta I, & i \neq j, \\ \big[H_i - (1 - p_i)\zeta I\big]/p_i, & i = j, \end{cases} \qquad b_{ij} = g_i - g_j.$$

All atomic Hessians are positive definite. Normalized row averages give (H_i, g_i) and normalized column averages give $(H_j, -g_j)$. The two partitions therefore have the same strictly convex atomic losses and the same union objective. Applying this lemma to the four-angle list completes the equality construction. The bound G and spectral set S constrain child gradients and Hessians; no corresponding atomic bounds are needed. Translating coordinates places the reference at any w_0.

G.3 Finite equal-mass realizations with controlled atomic curvature

Real-valued atomic weights are not essential to the ambiguity. The next corollary retains any desired fraction of the uniform diameter with an explicit number of equal-weight examples. It refines the realization of the classical recovery radius, not the recovery problem itself.

Corollary 4 (Finite equal-mass near-attainment)**.** *Let $a = \min S > 0$, $b = \max S$, and retain the information and radius $\mathcal{E}_r$ of Eq.* (53)*. For any integer $N \geq 1$, put $K = 4(N + r + 1)$ and $c_N = N/(N + r + 1)$. There are two K-child, equal-mass partitions of the same K^2 equal-weight strictly convex two-dimensional quadratic atoms whose endpoints admit the declared common transcript and satisfy*

$$\|P_L^1(w_0) - P_L^0(w_0)\| = 2c_N\mathcal{E}_r.$$

The two sides share their complete child-Hessian list, matrix spectral weights, and reference-point gradient distribution. The child spectra belong to S and their gradient norms are at most G. Every atomic gradient at w_0 has norm at most $2G$, and its Hessian eigenvalues lie in $[a, a + K(b - a)]$. If $b < 2a$, another realization bounds all atomic eigenvalues in $[2a - b, 2b - a]$, independently of N and K. The child Hessians are noncommuting whenever $\mathcal{E}_r > 0$.

Consequently, the supremum of separations over compatible same-pool pairs with the stated equal-mass and atomic bounds lies between $2c_N\mathcal{E}_r$ and $2\mathcal{E}_r$. Choosing $N \geq (r + 1)/\delta$ retains at least a

fraction $1-\delta$ for $0<\delta<1$. The pool and its common objective can be chosen by the construction; the corollary does not characterize a prescribed dataset or fix one objective across all depth/noise settings. The second curvature interval requires $b<2a$; its condition-number bound is not uniform as b/a approaches two.

Proof. Use coordinates $x=w-w_0$ and record displacements from w_0. Take a maximizing signed measure in Eq. (54). Represent its normalization by G as the signed marginal of a probability measure on $S\times\{-1,1\}$, adding cancelling mass if needed. Its moment vector consists of

$$\sigma\big(1,\phi_2(h),\dots,\phi_r(h),\phi_L(h)\big)\in\mathbb{R}^{r+1}.$$

Carathéodory's theorem preserves that vector using at most $r+2$ signed points with positive masses. The first coordinate is zero, so the positive- and negative-sign masses are each $1/2$. Couple their doubled probability marginals greedily. There are at most $r+1$ positive transport edges, with masses γ_e summing to one and paired values $(h_+^e,h_-^e)\in S^2$. If ν denotes the signed measure after this finite reduction, then

$$\frac{G}{2}\sum_e \gamma_e[f(h_+^e)-f(h_-^e)]=\int f\,d\nu$$

for every scalar f; ν still has the maximizing target moment and the same feasible observed moments.

Set $M_0=N+r+1$, $m_e=\lceil N\gamma_e\rceil$ and $t_e=GN\gamma_e/m_e\le G$. Replicate the four-angle block from the preceding subsection m_e times, using gradients $\pm t_eS_\theta e_1$. Give every child mass $1/(4M_0)$. Since $\sum_e m_e\le N+r+1$, fill the remaining $M_0-\sum_e m_e$ blocks with four children having Hessian h_0I and zero gradient, for any $h_0\in S$. There are exactly $K=4M_0$ children, independently of rounding outcomes. The amplitude compensates for the integer multiplicity: four-angle averaging yields

$$\sum_i \frac{1}{K}f(H_i)g_i^{\pm}=\pm\frac{N}{M_0}\left(\int f\,d\nu\right)e_1.$$

For $f=\phi_m$, the zero transcript has error at most $c_N\varepsilon_m$ on either side; the target separation is $2c_N\mathcal{E}_r$. For $f=1$ the mean gradient is zero. Each four-angle block has a sign-symmetric gradient law, including the padded blocks. The Hessian list and masses do not change between the two partitions. A nonzero target moment requires an edge with $h_+^e\neq h_-^e$, whose zero- and $\pi/4$-angle Hessians do not commute.

At $w_0=0$, use atoms with gradients g_i-g_j, equal weights $1/K^2$, and Hessians aI for $i\neq j$ and $KH_i-(K-1)aI$ for $i=j$. Their row and column means give (H_i,g_i) and $(H_j,-g_j)$. The atomic eigenvalues lie in $[a,a+K(b-a)]$ and the gradient norms are at most $2G$. For $b<2a$, alternatively take

$$A_{ij}=H_i+H_j-\bar H,\qquad \bar H=K^{-1}\sum_i H_i.$$

The same row/column means hold, and $(2a-b)I\preceq A_{ij}\preceq(2b-a)I$ is positive definite. Thus both are strictly convex same-pool realizations. A common constant $2G^2/a$ in the first construction, or $2G^2/(2a-b)$ in the second, makes every atom nonnegative if desired. Translating coordinates gives any reference input w_0. The indexed equal-weight examples may repeat. □

The finite budget is conservative; the next two-point result attains the full diameter with only sixteen atoms. Exact rational examples in the numerical supplement check compensated moments, integer padding, gradient-law matching, and the two atomic-curvature constructions, including three- and four-point cases retaining 90% of the sharp diameter.

G.4 Noise, uncertainty, and required test depth

For an endpoint tolerance $\tau>0$, the smallest depth that uniformly forces every compatible pair to be τ-close is

$$r_\star=\min\{1\le r<L:2\mathcal{E}_r\le\tau\},$$

with no early guarantee if this set is empty. Adding observations cannot increase $\mathcal{E}_r$, since their coefficients may be zero. Enlarging S or increasing an error budget cannot decrease it. Thus the criterion allows certified spectral clusters without treating their nominal centers as an exact spectrum.

For $S = \{a, b\}$ with $0 < a < b$, define $d_m = |\phi_m(a) - \phi_m(b)|$. Every zero-mass measure is $u(\delta_a - \delta_b)$, hence

$$\mathcal{E}_r = d_L \min\left\{G/2, \min_{2\le m\le r} \frac{\varepsilon_m}{d_m}\right\}. \tag{56}$$

Zero denominators impose no constraint; the inner minimum is $+\infty$ when $r = 1$. Under $0 < \eta b \le 1$, d_m increases with m. For equal noise radii ε and prior diameter $Gd_L > \tau$, a successful depth $r \ge 2$ is exactly one satisfying $d_r \ge 2\varepsilon d_L/\tau$. Even a two-point spectrum can therefore require more than two noisy observations for a useful tolerance.

A bounded equal-mass realization. The diameter in Eq. (56) is attained with four children and 16 equal-mass atoms. Put

$$g_\star = \min\left\{G, \min_{\substack{2\le m\le r\\ d_m>0}} \frac{2\varepsilon_m}{d_m}\right\}, \qquad c = \frac{a+b}{2}, \qquad \Delta = b - a,$$

with the same empty-minimum convention. At $w_0 = 0$, take children of mass $1/4$ with

$$H_1 = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad H_2 = \begin{pmatrix} c & -\Delta/2 \\ -\Delta/2 & c \end{pmatrix},$$
$$H_3 = \begin{pmatrix} b & 0 \\ 0 & a \end{pmatrix}, \quad H_4 = \begin{pmatrix} c & \Delta/2 \\ \Delta/2 & c \end{pmatrix},$$

and $(g_1, g_2, g_3, g_4) = g_\star(e_1, e_2, -e_1, -e_2)$. The two sides use (H_i, g_i) and $(H_i, -g_i)$, respectively. They share the full Hessian list and reference-gradient law, with $W_a = W_b = I/2$; H_1 and H_2 do not commute. Define atoms of mass $1/16$ by

$$\ell_{ij}(w) = \tfrac{1}{2} w^T A_{ij} w + (g_i - g_j)^T w + C,$$
$$A_{ij} = \begin{cases} 4H_i - 3aI, & i = j, \\ aI, & i \ne j. \end{cases} \tag{57}$$

Their Hessian eigenvalues lie in $[a, 4b - 3a]$. Normalized row and column averages give (H_i, g_i) and $(H_j, -g_j)$, so both partitions have the fixed union objective $F(w) = (a + b)\|w\|^2/4 + C$. One may set $C = 0$, or $C = 2G^2/a$ to make every atom nonnegative: $A_{ij} \succeq aI$ and $\|g_i - g_j\| \le 2G$ imply $\min_w \ell_{ij}(w) \ge C - 2G^2/a$.

The four-angle identity above, with G replaced by $g_\star$, gives

$$P_m^{\pm}(0) = \pm\frac{g_\star}{2}[\phi_m(a) - \phi_m(b)]e_1,$$
$$\|P_L^+(0) - P_L^-(0)\| = g_\star d_L = 2\mathcal{E}_r. \tag{58}$$

Both first endpoints vanish. The zero transcript has error $g_\star d_m/2 \le \varepsilon_m$ at every observed horizon, including the degenerate case $g_\star = 0$. The objective and atomic Hessians stay fixed across this family; the atomic linear terms may depend on the depth and error budgets. Sharpness is uniform over the family, not a diameter for every supplied pool or one pair attaining all depths.

For noiseless observations on $[a, b]$ and $L = r + 1$, monic Chebyshev approximation gives

$$\mathcal{E}_r = G|\eta|^{r+1}(b - a)^r/2^{2r-1}.$$

The diameter $2\mathcal{E}_r$ is precisely B_{r+1} from Theorem 2. With finite S of size q, noiseless depth $r \ge q$ gives zero radius; with an interval in S and $r < L$, the radius is positive. Substituting fixed interpolation coefficients into Eq. (53) gives a feasible upper bound, whereas the optimum trades approximation bias against noise amplification.

For different recorded prefixes, if their distance is at most d_m^{obs} and separate error budgets are $\varepsilon_m^0, \varepsilon_m^1$, replace ε_m by $(d_m^{\mathrm{obs}} + \varepsilon_m^0 + \varepsilon_m^1)/2$ in the upper bound. The uniform diameter is not the interval for a particular favorable transcript. Nor does the antipodal construction prove difficulty of a binary test on an absolute gap: its two target norms coincide. The result concerns endpoint identification. A single-map forecast additionally needs the aggregate gradient for the free constant, or a bound on its error.

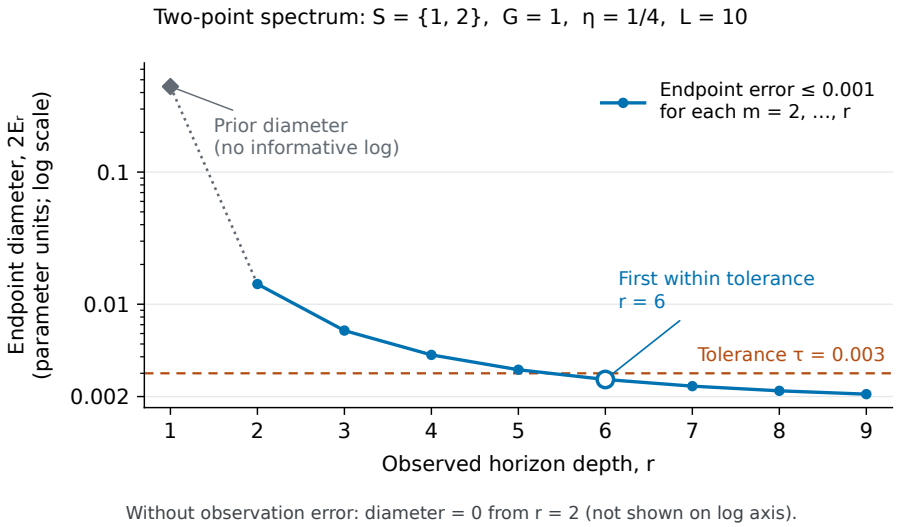

Figure 4: Noise changes the required observation depth even on a two-point spectrum. The curve is the exact worst-case endpoint diameter from Eq. (56), not an empirical trajectory. For the displayed fixed parameters, $2\mathcal{E}_5 = .0031918$ and $2\mathcal{E}_6 = .0026933$, so depth six is the first to meet tolerance .003. Without observation error, depth two identifies the endpoint exactly. Each point admits the 16-equal-atom realization above with atomic eigenvalues in $[1, 5]$. All curve values were computed as rational numbers.

H From exact identification to perturbed endpoint prediction

This section applies classical interpolation and its remainder formula (NIST, 2026b, Section 3.3(i)) to the quadratic endpoint response. It separates observation error from uncertainty about the curvature support. The first bound allows noncommuting symmetric quadratic Hessians. The subsequent coefficient and input-transfer calculations specialize to scalar quadratics. The reference input w_0 and nonzero step size η are fixed.

H.1 Observation noise and spectral-support error

For a single decomposition, let $y_m = P_m(w_0; \eta) - w_0$ and observe $\widetilde{y}_m$ with $\|\widetilde{y}_m - y_m\| \le \varepsilon_m$, $m = 1, \dots, q$. Distinct positive nominal eigenvalues $\widehat{h}_1, \dots, \widehat{h}_q$ give interpolation coefficients $a_{L,m}$ for horizon $L > q$. If actual child spectra lie in a compact positive interval I containing those nodes and $\|g_k(w_0)\| \le G$, then

$$\left\| y_L - \sum_{m=1}^{q} a_{L,m} \widetilde{y}_m \right\| \le \sum_{m=1}^{q} |a_{L,m}| \varepsilon_m + G K_{L,q}(\rho) \max_{k, \lambda \in \mathrm{spec}(H_k)} \prod_{j=1}^{q} |\lambda - \widehat{h}_j|, \tag{59}$$

where $\rho = \max_{h \in I} |1 - \eta h|$ and $K_{L,q}(\rho) = |\eta|^{q+1} \sum_{t=q}^{L-1} \binom{t}{q} \rho^{t-q}$. The proof applies the interpolation remainder to each child Hessian before averaging (Appendix F). For nominal support $(1, 3/2, 2)$ and $\eta = 1/4$, $y_4 = (89/64)y_1 - (97/32)y_2 + (23/8)y_3$ on exact support. Uniform observation error is amplified by exactly $467/64$ for this predictor. This separates identifiability from the conditioning of a Vandermonde reconstruction (Gautschi, 1990).

H.2 Predicting a later endpoint on known support

For distinct positive curvatures $h_1, \dots, h_q$, write p_j for their total masses and g_j for the conditional mean gradients at w_0. Set $r_j = 1 - \eta h_j$ and

$$y_E = P_E(w_0; \eta) - w_0 = \sum_j \frac{p_j g_j}{h_j} (r_j^E - 1), \qquad y_0 = 0.$$

Define $C(z) = (z-1)\prod_{j=1}^{q}(z-r_j)$, and let $\sum_{m=0}^{q} a_{L,m} z^m$ be the polynomial remainder of z^L on division by $C(z)$, for an integer $L > q$. Its values at $1, r_1, \ldots, r_q$ equal those of z^L, giving

$$y_L = \sum_{m=1}^{q} a_{L,m} y_m. \tag{60}$$

Indeed, interpolation gives $r_j^L - 1 = \sum_{m=0}^{q} a_{L,m}(r_j^m - 1)$; multiply by $p_j g_j / h_j$ and sum. Only the support locations, not the weights, are needed to compute these coefficients.

If $|\widetilde{y}_m - y_m| \le \varepsilon_m$, the predictor $\widehat{y}_L = \sum_{m=1}^{q} a_{L,m}\widetilde{y}_m$ satisfies

$$|\widehat{y}_L - y_L| \le \sum_{m=1}^{q} |a_{L,m}|\varepsilon_m. \tag{61}$$

Over the specified deterministic error box, this predictor's worst-case error is exactly the right-hand side, attained by sign-aligned errors. Polynomial division or the recurrence with characteristic polynomial C computes the coefficients directly from the support. This guarantee assumes exact coefficients and arithmetic; the coefficient norm quantifies observation sensitivity, while floating-point coefficient computation is a separate numerical question.

H.3 Approximate curvature support

Now form $a_{L,m}$ from nominal distinct positive curvatures $\widehat{h}_1, \ldots, \widehat{h}_q$. The actual curvature law μ may have any number of support points in a compact positive interval I containing both supports. Assume its conditional input gradient obeys $|g(h)| \le G$ almost everywhere. Define

$$\phi_E(h) = -\eta \sum_{t=0}^{E-1} (1-\eta h)^t, \qquad J_L(h) = \sum_{m=1}^{q} a_{L,m}\phi_m(h).$$

The polynomial J_L has degree at most $q-1$ and agrees with ϕ_L at each nominal node, by (60). It is therefore the Lagrange interpolant of ϕ_L. With $\rho = \max_{h\in I}|1-\eta h|$, the standard interpolation remainder and direct differentiation give

$$|\phi_L(h) - J_L(h)| \le K_{L,q}(\rho)\prod_{j=1}^{q}|h-\widehat{h}_j|, \tag{62}$$

$$K_{L,q}(\rho) = |\eta|^{q+1}\sum_{t=q}^{L-1}\binom{t}{q}\rho^{t-q}, \qquad \frac{\phi_L^{(q)}(h)}{q!} = (-\eta)^{q+1}\sum_{t=q}^{L-1}\binom{t}{q}(1-\eta h)^{t-q}.$$

Integrating $g(h)[\phi_L(h) - J_L(h)]$ against μ proves (59), after adding the observation-error term. This proof permits an unknown number of actual curvature values and does not require a bijection between actual and nominal nodes.

For an explicit proximity budget, assign each actual curvature to a nominal node j with $|h - \widehat{h}_j| \le \delta_j$. Then

$$\prod_k |h-\widehat{h}_k| \le W_j := \delta_j \prod_{k\ne j}(|\widehat{h}_j - \widehat{h}_k| + \delta_j).$$

The support-error term is at most $GK_{L,q}(\rho)\max_j W_j$. Under a common radius δ, a simpler bound is $GK_{L,q}(\rho)\delta(\Delta_h + \delta)^{q-1}$, where $\Delta_h = \max_j \widehat{h}_j - \min_j \widehat{h}_j$. For $\rho \le 1$, $K_{L,q}(\rho) \le |\eta|^{q+1}\binom{L}{q+1}$. These are absolute bounds; support proximity need not ensure small relative error when the true endpoint displacement nearly cancels.

H.4 Changing the input also requires slope information

Let $S_L = \int (1-\eta h)^L \, d\mu(h)$. The actual map at another input is $P_L(w) = w_0 + y_L + (w - w_0)S_L$. For nominal masses $\widehat{p}_j$, let $\widehat{S}_L = \sum_j \widehat{p}_j(1-\eta\widehat{h}_j)^L$. Write π_j for the actual mass assigned to nominal node j by the same proximity rule, set $\delta_{\max} = \max_j \delta_j$, and assume $\sum_j |\pi_j - \widehat{p}_j| \le \kappa$. The mean-value theorem gives

$$|S_L - \widehat{S}_L| \le L|\eta|\rho^{L-1}\delta_{\max} + \kappa\rho^L.$$

Hence a predictor using the nominal slope has the bound

$$|\widehat{P}_L(w) - P_L(w)| \le \sum_{m=1}^{q} |a_{L,m}|\varepsilon_m + GK_{L,q}(\rho)\max_j W_j + |w - w_0|\big(L|\eta|\rho^{L-1}\delta_{\max} + \kappa\rho^L\big). \tag{63}$$

The mass discrepancy here is the full ℓ_1 discrepancy, not half of it. Weight uncertainty alone contributes nothing at w_0 if support is exact, but may change the slope at other inputs.

Two different laws. For two scalar decompositions, let q count the union of their distinct positive curvatures. Their displacement difference is annihilated by $(z-1)\prod_{h\in\mathcal{H}_{\text{union}}}(z-(1-\eta h))$. It is zero at horizon zero. Equality at horizons $1,\dots,q$ therefore implies equality at all later horizons at the same reference input. The equivalent moment argument identifies the weighted coefficients $p^b(h)g^b(h)$, with zero coefficients on absent support, so the conclusion also holds at every step size there. It does not imply equality at other inputs: $F^0(w) = w^2/2$ and $F^1(w) = w^2$ both stay at zero from $w_0 = 0$, but their maps at $\eta = 1/4$ are $(3/4)^E w$ and $(1/2)^E w$. The common weighted law in Theorem 2 supplies the additional slope equality and its stronger all-input conclusion.

H.5 An exact three-curvature example

Take $w_0 = 0$, $\eta = 1/4$, $h = (1, 3/2, 2)$ and $p = (1/4, 1/2, 1/4)$. Then

$$C(z) = z^4 - \frac{23}{8}z^3 + \frac{97}{32}z^2 - \frac{89}{64}z + \frac{15}{64}, \quad y_4 = \frac{89}{64}y_1 - \frac{97}{32}y_2 + \frac{23}{8}y_3.$$

The noise amplification for uniform observation error is exactly $467/64 = 7.296875$, and $S_4 = 1401/8192$. For gradients $(G, -G, G)$, the observations are $y_1 = y_2 = 0$, $y_3 = -G/512$, and $y_4 = -23G/4096$. The support interpolation residual factors exactly as

$$\phi_4(h) - J_4(h) = \frac{1}{256}(h-1)(h-3/2)(h-2).$$

If actual curvatures lie within $\delta < 1/4$ of the nominal nodes, the reference-input support error is at most $G\delta(1/2+\delta)(1+\delta)/256$. The supplement verifies these identities with exact rational arithmetic.

I COMPARISON WITH ADJACENT LOCAL-UPDATE ANALYSES

Table 5: How the present identifiability question differs from adjacent local-update analyses. The third column refers only to the matched pair in Theorem 1.

Work	Fixed or varied object	Matched-pair result	Primary output
Local SGD / client drift (Karimireddy et al., 2020; Gorbunov et al., 2021; Glasgow et al., 2022)	Fix the decomposition; vary local-step behavior, heterogeneity assumptions, or correction	Not targeted: client objectives remain attached to fixed clients	Drift or convergence bounds; corrected updates
Surrogate objectives / iterated maps (Charles & Konečný, 2021; Charles & Rush, 2022)	Fix the decomposition; characterize its composed local map or surrogate	Not targeted across alternative decompositions	Surrogate losses, fixed points, or conservatism
Objective / fixed-point consistency (Wang et al., 2020; Malinovskiy et al., 2020; Wang et al., 2021)	Change local work, effective weights, or operators; ask which objective or equilibrium results	Here work, masses, objective, and fixed point may agree while endpoints differ	Limiting-objective or fixed-point consistency
Hierarchy / sample topology (Wang et al., 2022; Huang et al., 2022)	Specify or design grouping or topology; analyze consensus or convergence	At w_0, upward/downward gradient dispersions can agree while D_2 differs	Rates or topology frameworks
Gradient / Hessian heterogeneity (Karimireddy et al., 2020; Wang et al., 2024; Patel et al., 2025; Luo et al., 2025)	Fix client objectives; bound averaged-local/central mismatch, Hessian similarity, or consensus error	For quadratics the mismatch contains $A_{\mathcal{P}}$; its unpaired marginals do not fix the sign	Convergence or consensus bounds
This work	Fix atoms, objective, mass, input, rule, horizon, and work; vary only grouping	Pairing distinguishes the exact $E = 2$ construction; explicit remainders bound the higher-order error	Exact nonidentifiability and finite-step deviation bounds

Convergence, correction, and communication. FedAvg combines local training with parameter averaging (McMahan et al., 2017). Local-SGD analyses quantify the resulting communication–convergence trade-off (Stich, 2019; Haddadpour et al., 2019); non-IID analyses and comparisons with minibatch SGD characterize the role of heterogeneity (Li et al., 2020b; Woodworth et al., 2020). FedProx modifies local objectives with a proximal term (Li et al., 2020a), while SCAFFOLD corrects client drift (Karimireddy et al., 2020). Multi-level local SGD studies interactions between local, sub-network, and global communication (Castiglia et al., 2021). These results motivate the local-update map studied here. We keep the update rule and work fixed and instead ask what information distinguishes alternative assignments of the same losses.

Second-order heterogeneity. Recent analyses sharpen Local-SGD and SCAFFOLD convergence guarantees (Luo et al., 2025), refine consensus-error bounds using second-order heterogeneity (Patel et al., 2025), and extend such convergence guarantees to general convex objectives with nearly matching lower bounds (Patel et al., 2026). Our lower bounds have a different target: ambiguity between finite-step maps compatible with specified information, rather than optimization error relative to a minimizer. The matched-constant calculation below makes this distinction explicit for two nearby bounds.

I.1 THE KNOWN QUADRATIC MAP AND THE IDENTIFICATION QUESTION

Charles & Konečný (2021, Eq. (6), Theorem 1) define a client distortion matrix $Q_i = \sum_{t=0}^{E-1}(I - \eta H_i)^t$ when their proximal parameter is zero and their gradient weights are all one. Their Lemma 7, Eq. (21), gives the gradient sequence $(I - \eta H_i)^t g_i(w)$; their Lemma 9, Eqs. (29)–(30), gives its spectral filter. Averaging and using server step size η yields

$$P_E(w) = w - \eta \sum_i p_i Q_i g_i(w) = w + \sum_{r=0}^{E-1} \binom{E}{r+1} (-\eta)^{r+1} \sum_i p_i H_i^r g_i(w).$$

Thus the polynomial in Proposition 5 is a direct expansion of their forward-map formula, not a new update formula. Their surrogate-loss characterization already determines the whole map when the child objectives are given. Our question instead specifies information shared by different assignments: unpaired derivatives or a finite sequence of endpoint observations. The new bounds quantify the remaining ambiguity under those constraints and realize the extremal cases with partitions of the same atomic losses.

I.2 FINITE MOMENTS AND CLASSICAL RECONSTRUCTION

Prony-type methods recover finite-support measures from moments (Kunis et al., 2016, Section 1). The coordinatewise correspondence here is $f_\ell(r) = (M_r)_\ell = \sum_\lambda \lambda^r (R_\lambda)_\ell$. If $p(z) = \prod_\lambda (z - \lambda) = \sum_{j=0}^q c_j z^j$, then $\sum_{j=0}^q c_j M_{r+j} = 0$: the same annihilating-polynomial recurrence applies. With supplied support, recovering the residues is a Vandermonde solve; recovering unknown support is a different inverse problem. Our first-q equivalence concerns equality on the actual union support, not its reconstruction from q observations. The extra constraint in our extremal problem is that both measures must arise from partitions of one strictly convex atomic loss pool, with matched objective, derivative laws and, where specified, matrix spectral weights. The same-atom attainments establish sharpness inside this constrained class.

Classical Vandermonde analysis separates invertibility from numerical conditioning (Gautschi, 1990); Chebyshev approximation supplies extremal polynomial and interpolation tools (Trefethen, 2013). Computational optimal recovery studies the best reconstruction from limited, possibly corrupted observations (Ettehad & Foucart, 2020; Foucart, 2021). Here those tools bound the ambiguity of local-update observations, while the same-loss constructions establish attainability within the optimization model. Appendix H quantifies observation-noise amplification and spectral-support error separately.

I.3 HETEROGENEITY CONSTANTS ON THE MATCHED CONSTRUCTION

Theorem 1 preserves specific constants from two nearby analyses, even though the realized gaps differ. The evaluations below are our calculations on this witness, not claims that the cited works studied these alternative partitions. Write $a = \sqrt{v_g}$, $b = \sqrt{v_H}$, and $x = w - w_0$. The row and column children in Appendix A.4 have gradients at zero $\bar{g} + as_k$ and Hessians $\bar{H} + bu_k$, where $u = s$ for rows and $u = t$ for columns. The centered vectors satisfy $\|s\|^2 = \|t\|^2 = K$, $s^\top t = 0$, and t is a permutation of s.

Wang et al. (2024, Proposition 5.3) give the quadratic choice $L_h = \max_k \|H_k - \bar{H}\|_{\mathrm{op}} = b\|u\|_\infty$, identical for the two partitions. Their Eq. (9) bounds the squared averaged-gradient discrepancy by $L_h^2 K^{-1} \sum_k e_k^2$, where $e_k = z_k - \bar{z}$ and $\sum_k e_k = 0$. On this scalar construction the discrepancy is $bK^{-1} \sum_k u_k e_k$. Cauchy–Schwarz bounds its square by $b^2 K^{-1} \sum_k e_k^2$, with equality at $e_k = cu_k$. Hence the smallest admissible L_h is also the same, namely b. This last equality is our calculation for the witness, not a general matrix-valued formula from the cited proposition.

At the first local iterates, $e_k = -\eta a s_k$ in both partitions. Their mean squared spread is the same, $\eta^2 v_g$, while the signed gradient discrepancies are $-\eta ab$ and zero. Thus even the optimal right-hand side in Eq. (9) is identical although one partition attains it and the other cancels. Multiplication by $-\eta$ yields the exact two-step gaps in Theorem 1.

The smallest Hessian-difference parameter in Patel et al. (2025, Assumption 6) is likewise identical:

$$\tau = \sup_{k,\ell,x} |H_k(x) - H_\ell(x)| = b(\max s - \min s) = b(\max t - \min t).$$

For Figure 1's sign-pattern example, $b = .7$, so both the stated and optimal L_h equal .7, and $\tau = 1.4$. This matches selected heterogeneity constants, not every assumption or parameter of the cited convergence results. For example, the two gradient dispersion functions are $(a + bx)^2$ and $a^2 + b^2x^2$, and child-minimizer distances need not agree. The comparisons distinguish valid worst-case bounds from identification of the realized update; they do not contradict either bound. Theorem 2 separately shows that even the entire dispersion function can agree when the distinguishing moment occurs at a later horizon.

J Evaluation protocols

J.1 Synchronization and ownership

The analyzed system is a synchronized, owner-coordinated hierarchy. The root is the unit of logical data ownership, participation, work, and server mass. Its children are disjoint shards that co-participate, receive the same input, run the same stateless local map, and expose no independent sampling decision. The root coordinator applies the prescribed masses and preaggregates one root update; a failed child invalidates that contribution. If children are visible to the server, the system must still enforce owner-level co-participation and mass equivalence. Candidate partitions must already satisfy privacy, consent, data-residency, fairness, and availability requirements. Sensitive gradients, HVPs, and partition metadata should remain inside the owner's trusted boundary; the certificate supplies no privacy mechanism for releasing them.

J.2 Exact root-to-child construction

For FEMNIST and CIFAR roots, each example has immutable pair $(\texttt{client_id}, \texttt{rowid})$. A fixed, label-blind permutation of these identifiers orders the examples; zero-based rank j is assigned to child $j \bmod K$. Shakespeare applies the same procedure to immutable source-snippet IDs before converting each snippet into byte-level context/target pairs, so all examples from a snippet remain in one child. Child weights are $p_{rk} = |D_{rk}|/|D_r|$. Labels and outcomes never enter this construction. Current prediction-cohort memberships are supplied in the numerical supplement; historical task-model records remain local.

The structured-split evaluation instead constructs twelve candidate partitions at fixed $K = 4$. For variant $v \in \{0, 1, 2, 3\}$, the hash family sorts examples by a frozen salted digest and assigns strided ranks to children; the label-block family sorts by (label, frozen salted digest) and takes four contiguous near-equal chunks; and source-order sorts by immutable row ID, rotates by $v \max(1, \lfloor N/4 \rfloor)$, and takes four contiguous near-equal chunks. Source order is an acquisition-order proxy, not documented time. Each candidate contains every root example exactly once and has at least four examples per child. Candidate manifests are frozen before endpoint evaluation.

J.3 Data, models, and independent units

FEMNIST and Shakespeare are from the LEAF benchmark (Caldas et al., 2018); The CIFAR dataset is described by Krizhevsky (2009). The frozen-feature encoder uses ImageNet-pretrained residual-network features (He et al., 2016; Deng et al., 2009). The from-scratch CNN uses Group Normalization (Wu & He, 2018).

Table 6: Evaluation regimes. Smooth evaluations use the logical root as the independent unit; task-model comparisons use paired seeds. Repeated map evaluations are not independent observations.

Regime	Model/objective	Protocol and independent unit
FEMNIST task model	Small writer CNN	128 writer roots, 40 rounds, 8 roots/round; seeds 607/619/631/647/661/677; factors 1/2/4/8 and $E = 1/2/5/10$; root schedules are paired.
Shakespeare task model	8-byte-context feed-forward MLP (32-dimensional embedding; 128-unit ReLU hidden layer)	128 speaking-role roots, 30 rounds, 8 roots/round; seeds 701/719/733/751/769/787; the same factor and local-step grid; root schedules are paired.
CIFAR-100 from scratch	Three-block 32/64/128-channel GroupNorm CNN	100 class-balanced roots of 500 examples; 60 rounds; seeds 809/823/839/853/877/881; factors 1/2/4/8 and $E = 1/5/10$.
CIFAR-100 features	Frozen ImageNet ResNet-18 plus linear head	512-dimensional post-pooling features; 100 rounds; seeds 907/919/937/953/971/983; factors 1/8 and $E = 1/5/10$; no encoder fine-tuning.
Smooth certificate	Unregularized binary logistic objective in float64	128 FEMNIST roots at $K = 8$; $E = 1/2/5/10$ and $\eta = .05/.025/.0125/.00625$; 2,048 maps plus 256 independent extended-precision checks.
Candidate ranking	Same smooth logistic objective	64 roots, 12 pre-specified $K = 4$ candidates/root, 12 (E, η) settings; root is the bootstrap unit.

The task-model learning rate is 0.05 for the two natural-user regimes and CIFAR from scratch, and 0.1 for the frozen-feature head; weight decay is 10^{-4} and task-model momentum is zero unless the optimizer-boundary experiment explicitly states otherwise. The CIFAR-100 official test set is split within class into 4,000 temperature-fitting (Guo et al., 2017) and 6,000 untouched evaluation examples.

J.4 Task-model endpoint-gap estimand

For seed s, let $\theta_{s,T}^{(K)}$ and $\theta_{s,T}^{(1)}$ denote the saved final model states of the factor-K and union arms after the same T training rounds. All saved state tensors are flattened in matching name order and concatenated in float64. The reported relative gap is the single global Euclidean ratio

$$d_s^{(K)} = \frac{\|\theta_{s,T}^{(K)} - \theta_{s,T}^{(1)}\|_2}{\max(\|\theta_{s,T}^{(1)}\|_2, 10^{-15})}. \tag{64}$$

The task-model gap figures report the median of these six paired seed values at each factor and local horizon; they do not average gaps over rounds. The earlier held-out task-effect summary instead reports the mean and two-sided 95% Student-t interval across its six held-out seed values. Its task-metric columns use paired factor-8 minus union differences in final evaluation metrics. These final-trajectory summaries differ from both the same-input smooth-map gaps and the round-median batching control in the earlier batching comparison.

J.5 Certificate and ranking estimands

For every example, the smooth objective uses

$$\ell_i(w) = \log\big(1 + \exp(-y_i x_i^\top w)\big), \qquad x_i = [\mathrm{vec}(\mathrm{pixel}_i), 1]/\sqrt{785}, \quad y_i \in \{-1, +1\}, \tag{65}$$

where parity maps the FEMNIST class to the binary label and the common reference input is $w = 0 \in \mathbb{R}^{785}$. For any nonempty subset S, the implementation computes the following analytic global bounds and rounds each upward by one float64 representable value:

$$G_S = \frac{1}{|S|}\sum_{i\in S}\|x_i\|, \qquad L_S = \frac{1}{4|S|}\sum_{i\in S}\|x_i\|^2, \qquad \rho_S = \frac{1}{6\sqrt{3}|S|}\sum_{i\in S}\|x_i\|^3. \tag{66}$$

The scalar logistic derivative bounds make these inequalities valid for every $w \in \mathbb{R}^{785}$; hence $\mathcal{C} = \mathbb{R}^{785}$ satisfies the containment condition. The constants depend only on the fixed feature matrix, not on an observed endpoint or defect. Across the 128 union objectives in the direct smooth evaluation, G_S ranges from 0.103652 to 0.207546, L_S from 0.002779 to 0.011035, and ρ_S from 0.000119 to 0.000920. Child and union terms in every reported $R_{\mathcal{P},E}$ use the same subset-specific formula (66).

The inequalities above are analytic real-arithmetic certificates. The tables evaluate their terms in float64: the feature-norm bounds are rounded upward by one representable value, dense symmetric

eigensolvers estimate centered-Hessian norms, and coverage comparisons allow 10^{-12} absolute plus 10^{-10} relative slack. Because the full computation is not outward-rounded interval arithmetic, the empirical coverage and decision counts are numerical evaluations of the analytic certificate, not a machine proof of each real-valued instance.

For the smooth finite-map study, both child and union objectives are evaluated from the same input. Full dense symmetric float64 eigendecomposition supplies centered-Hessian operator norms. The evaluation computes $A_{\mathcal{P}}$, V_g, V_H, $C_{\mathcal{P}}$, B^A, and B^C.

For candidate ranking, each root and each of the twelve (E, η) settings first yields a Spearman correlation across twelve candidate partitions. A root score is the median of its twelve setting scores. Final means and medians aggregate 64 roots. Percentile intervals for a mean root score resample roots 100,000 times with seed 20260804; they are not intervals for the reported median. A common bootstrap draw matrix is used across ranking scores. The analysis plan designates $C_{\mathcal{P}}$ as the confirmatory ranking score; $\|A_{\mathcal{P}}\|$ is a secondary analysis of the same recorded maps.

K Supporting smooth-objective evaluations

K.1 Numerical evaluation of the analytic scalar certificate

The 2,048 records are 128 roots times sixteen same-input (E, η) settings: 512 affine $E = 1$ controls and 1,536 nontrivial $E > 1$ finite maps. They are repeated deterministic map evaluations, not independent observations. Under the float64 procedure in Appendix J.5, both B^A and B^C pass the numerical coverage comparison for all 2,048 records. The maximum cross-check error against the earlier A-based calculation is 6.66×10^{-16}. These counts evaluate an analytic certificate; they are not outward-rounded machine proofs of the individual real-valued cases.

Table 7: Certificate-to-gap ratios in the smooth evaluation.

E	median $B^A/\|D\|$	median $B^C/\|D\|$	p95 $B^C/\|D\|$	max.
2	1.024	2.747	3.428	4.517
5	1.103	2.826	3.577	4.770
10	1.236	2.980	4.051	5.192

Median $C_{\mathcal{P}}/\|A_{\mathcal{P}}\|$ is 2.704 and its 95th percentile is 3.379. These values measure the Cauchy–Schwarz relaxation in this smooth setting; they do not imply comparable tightness for task networks.

The propagation evaluation supplies an additional nonzero-input check for the $A_{\mathcal{P}}$ interval. Six union trajectories use $E = 5$, $\eta = .05$, and 40 rounds. Excluding their six zero initial inputs leaves 234 same-input maps at state norms ranging from 0.002515 to 0.101559 (median 0.051708). All 234/234 are covered. The minimum, median, 95th-percentile, and maximum values of $B^A/\|D\|$ are 1.311, 1.447, 1.573, and 1.654; the corresponding certified remainder share has median 30.86%, 95th percentile 36.41%, and maximum 39.55%. This check uses the same smooth logistic objective and reports only the $A_{\mathcal{P}}$ interval; it is not a certificate evaluation on the task networks.

A separate checkpoint-grid audit evaluates both B^A and B^C away from the origin, at the same $K = 8$. For each of the same 128 FEMNIST roots, union GD with learning rate .025 produces reference states after 0, 10, 50, and 100 steps. At every state the audit evaluates $E \in \{2, 5, 10\}$ and $\eta \in \{.00625, .0125, .025, .05\}$, for 6,144 finite maps in total. Both bounds pass the float64 numerical coverage comparison on all 1,536 maps at each checkpoint. Table 8 reports the median tightness ratios; the small change across checkpoints is evidence for state stability within this smooth objective, not a guarantee for task networks or arbitrary training trajectories.

Table 9: New correct tolerance passes out of eight ridge problems, relative to the common zero/one-log baseline. Here $\tau = .01$, aggregate coordinates are rounded to 10^{-8}, and $\eta = \alpha/\max_k \lambda_{\max}(H_k)$. The primary comparison is $(\alpha, L) = (.75, 6)$. Predictors here receive child covariances and the fitted head; the matrix column is classical recovery, not the spectrum-only radius of Theorem 3. All baseline intervals remain unresolved in these six settings.

α	L	Baseline	Cheb. product	Scalar spectrum	Matrix
0.25	4	0/8	8/8	8/8	8/8
0.25	6	0/8	8/8	8/8	8/8
0.25	10	0/8	0/8	8/8	8/8
0.75	4	0/8	0/8	8/8	8/8
0.75	6	0/8	0/8	0/8	8/8
0.75	10	0/8	0/8	0/8	8/8

Table 8: Conventional pooled median bound-to-gap ratios at union-GD states ($K = 8$). Each cell uses 512 maps (128 roots $\times$ four learning rates), averages the two central observations, and reports $B^A/\|D\|$ followed by $B^C/\|D\|$.

Checkpoint step	$E = 2$	$E = 5$	$E = 10$
0	1.0242 / 2.7472	1.1035 / 2.8257	1.2356 / 2.9799
10	1.0242 / 2.7474	1.1035 / 2.8257	1.2357 / 2.9800
50	1.0242 / 2.7480	1.1035 / 2.8260	1.2358 / 2.9804
100	1.0242 / 2.7483	1.1036 / 2.8264	1.2360 / 2.9810

L TOLERANCE DECISIONS FOR A FROZEN-FEATURE RIDGE HEAD

On the primary panel, three-log matrix recovery resolves all eight roots left unresolved by the reported covariance-plus-one-update baselines. Neither the original Chebyshev product bound nor the tighter scalar spectrum-residual suite resolves a root. All methods receive the same full child covariance matrices and fitted head.

L.1 INFORMATION AND FIXED LEARNING INSTANCE

Frozen 512-dimensional ResNet-18 features of CIFAR-100 receive one outcome-blind Gaussian orthonormal projection to $d = 64$ and a common feature-only scale bounding row norms by one. A fixed permutation of 50,000 training indices supplies eight disjoint 2,048-example roots, each with eight 256-example children and $p_k = 1/8$. All $c = 100$ one-hot classes are retained, without a bias coordinate:

$$F_k(W) = \frac{\|X_k W - Y_k\|_F^2}{2n_k} + \frac{\gamma}{2}\|W\|_F^2, \quad C_k = \frac{X_k^\top X_k}{n_k}, \quad H_k = C_k + \gamma I, \qquad \gamma = \frac{1}{1024}.$$

The known reference head W_0 is fitted to that root's union objective. Writing $B_k = X_k^\top Y_k/n_k$ and bars for mass-weighted means, its supplied fit contract is $\|\bar{H}W_0 - \bar{B}\|_F \le \delta_{\text{fit}}$, with $\delta_{\text{fit}} = 10^{-10}(1 + \|\bar{H}W_0\|_F)$. The floating-point solution is not treated as exactly stationary.

All predictors receive the same independently supplied full covariances, masses, fitted head and fit bound, plus C_{probe} from the 10,000 test feature rows without their labels. The zero-log baseline thus includes the label-dependent fitted head. Optional aggregate displacements $y_m = \sum_k p_k(W_{k,m} - W_0)$ are recorded for $m = 1, 2, 3$ at $\eta = \alpha/U$, where $U = \max_k \lambda_{\max}(H_k)$ and $\alpha \in \{.25, .75\}$. Predictors receive neither raw feature rows, labels, B_k, individual child endpoints nor target-horizon aggregates. The target is $D_L = \sum_k p_k W_{k,L} - W_{\text{union},L}$ for $L \in \{4, 6, 10\}$.

The supplied logs $\widetilde{y}_m$ are either unrounded float64 recordings or once-per-aggregate coordinate roundings with $\Delta = 10^{-8}$. With $r_{\text{round}} = \Delta\sqrt{dc}/2$ (zero for unrounded recordings), the public observation allowance is $\varepsilon_m = r_{\text{round}} + 10^{-10}(1 + \|\widetilde{y}_m\|_F + r_{\text{round}})$. The encodings are separate predictor inputs; both predictions precede target evaluation. This information interface provides no privacy mechanism.

L.2 Scalar forecast and classical matrix comparison

Scalar interpolation. Let $\phi_m(H) = -\eta\sum_{t=0}^{m-1}(I-\eta H)^t$. For $q = 1$ or 3, solve $\sum_{m=1}^{q} a_m\phi_m(h_j) = \phi_L(h_j)$ at the fixed nominal nodes. The forecast from Eq. (59), with the union displacement reconstructed from the first log, is unchanged:

$$\widehat{D}_L = \sum_{m=1}^{q} a_m\widetilde{y}_m + \phi_L(\bar{H})\widetilde{y}_1/\eta.$$

Primary nodes are Chebyshev roots on $[\gamma, U]$; the prespecified secondary rule uses Lloyd centers of public child eigenvalues, initialized at equally spaced empirical quantiles, with at most 100 iterations and a fixed convergence tolerance. Nodes and parameters are target-independent.

The independently available gradient bounds are $G_k = \|H_kW_0\|_F + \sqrt{\|C_k\|_{\mathrm{op}}}$ and $G = \max_k G_k$. We report separately the generic product remainder of Eq. (59) and the exact public-spectrum residual evaluation of the same proof:

$$B_{\mathrm{prod}} = GK_{L,q}(\rho)\max_{\lambda\in\cup_k \mathrm{spec}(H_k)}\prod_{j=1}^{q}|\lambda-h_j|,$$

$$B_{\mathrm{spec}} = \sum_k p_kG_k\max_{\lambda\in\mathrm{spec}(H_k)}\left|\phi_L(\lambda)-\sum_{m=1}^{q}a_m\phi_m(\lambda)\right|.$$

Each radius adds the same observation term

$$N = \|a_1I+\phi_L(\bar{H})/\eta\|_{\mathrm{op}}\varepsilon_1 + \sum_{m=2}^{q}|a_m|\varepsilon_m.$$

The combined coefficient accounts for reuse of $\widetilde{y}_1$. B_{prod} and B_{spec} bound the same forecast.

Matrix recovery. The one-hot target condition gives the relaxation $B_k = C_k^{1/2}V_k$, $\|V_k\|_F \le 1$. For each acquired log define

$$A_m = [p_k\phi_m(H_k)C_k^{1/2}]_k, \qquad z_m = \sum_k p_k\phi_m(H_k)H_kW_0 - \widetilde{y}_m.$$

All observation budgets also include $A_0 = [p_kC_k^{1/2}]_k$, $z_0 = \bar{H}W_0$ with error $\varepsilon_0 = \delta_{\mathrm{fit}}$. Writing V for the stacked V_k, the target is $d_L - T_LV$, where

$$d_L = \left[\sum_k p_k\phi_L(H_k)H_k - \phi_L(\bar{H})\bar{H}\right]W_0,$$
$$T_L = [p_k(\phi_L(H_k)-\phi_L(\bar{H}))C_k^{1/2}]_k.$$

For a linear recovery operator R, its center and radius are

$$\widehat{D}_L = d_L - Rz, \qquad \mathcal{R} = \sum_k\|(T_L-RA)_k\|_{\mathrm{op}} + \sum_m\|R_m\|_{\mathrm{op}}\varepsilon_m.$$

Here R_m is the original d-column observation block, including $m = 0$. Rows of A are normalized to form A_n before SVD, and recovery is mapped back to original units. We retain the zero operator, a pseudoinverse with relative cutoff 10^{-12}, and ridge ratios $\lambda/s_{\max}^2 \in \{10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$. This classical comparator is not globally optimal information recovery; the ellipsoid relaxation need not correspond to realizable labelings.

L.3 Tolerance decisions and complete results

The comparison metric is the RMS change of the complete 100-dimensional vector, $s_L = \|C_{\mathrm{probe}}^{1/2}D_L\|_F$. A candidate center with Frobenius radius $\mathcal{R}$ yields $[\max(0,\widehat{s}-r),\widehat{s}+r]$, where $\widehat{s} = \|C_{\mathrm{probe}}^{1/2}\widehat{D}_L\|_F$ and $r = \sqrt{\|C_{\mathrm{probe}}\|_{\mathrm{op}}}\mathcal{R}$. Pass means the upper endpoint is at most τ; fail means

the lower endpoint exceeds τ; otherwise the decision abstains. The fixed tolerances .01 and .05 are illustrative score-space thresholds, not validated utility thresholds.

The common baseline intersects every valid zero/one-log matrix and scalar spectrum-residual interval. Each richer suite retains this baseline and adds its three-log candidates, preserving decisions from larger-radius candidates. Intersections are enlarged by $10^{-10}(1 + \max(|\ell|, |u|, \tau))$ at each end; an inconsistent empty intersection is an integrity failure. Scalar and matrix additions are counted separately as new correct decisions from a baseline abstention.

Table 10: Every fixed panel; all entries are counts out of eight roots. Z and B are decisions already made by the zero-log and common zero/one-log baselines. Remaining columns are newly correct decisions after adding three-log candidates: C is Chebyshev, L is the secondary Lloyd rule, subscripts p and s denote product and exact-spectrum bounds, and M uses all matrix candidates. Every definite decision is a pass. The full scalar spectrum suite has the same counts as L_s here.

α	L	τ	Z	B	C_p	C_s	L_p	L_s	M
Unrounded aggregate logs									
0.25	4	0.01	0	0	8	8	8	8	8
0.25	4	0.05	8	8	0	0	0	0	0
0.25	6	0.01	0	0	8	8	8	8	8
0.25	6	0.05	8	8	0	0	0	0	0
0.25	10	0.01	0	0	0	0	0	8	8
0.25	10	0.05	0	0	8	8	8	8	8
0.75	4	0.01	0	0	0	0	5	8	8
0.75	4	0.05	0	0	8	8	8	8	8
0.75	6	0.01	0	0	0	0	0	0	8
0.75	6	0.05	0	0	0	0	0	8	8
0.75	10	0.01	0	0	0	0	0	0	8
0.75	10	0.05	0	0	0	0	0	0	8
Coordinate-rounded aggregate logs, $\Delta = 10^{-8}$									
0.25	4	0.01	0	0	8	8	8	8	8
0.25	4	0.05	8	8	0	0	0	0	0
0.25	6	0.01	0	0	8	8	8	8	8
0.25	6	0.05	8	8	0	0	0	0	0
0.25	10	0.01	0	0	0	0	0	8	8
0.25	10	0.05	0	0	8	8	8	8	8
0.75	4	0.01	0	0	0	0	5	8	8
0.75	4	0.05	0	0	8	8	8	8	8
0.75	6	0.01	0	0	0	0	0	0	8
0.75	6	0.05	0	0	0	0	0	8	8
0.75	10	0.01	0	0	0	0	0	0	8
0.75	10	0.05	0	0	0	0	0	0	8

The primary panel fixes $\alpha = .75$, $L = 6$, rounded logs and $\tau = .01$. Its matrix suite adds eight correct passes, while the scalar suite adds zero. The median true score change is 7.1234×10^{-4}; baseline upper endpoints range from .13133 to .14549, scalar spectrum-suite uppers from .022754 to .048555, and matrix uppers from 5.5527×10^{-4} to 8.4275×10^{-4}. The original Chebyshev product bound adds eight passes at the prespecified secondary settings $\alpha = .25$, $L = 4, 6$, $\tau = .01$. At $\alpha = .25, L = 10$, the scalar success instead requires the Lloyd exact-spectrum bound; at $\alpha = .75, L = 4$, that bound resolves eight roots while the Lloyd product bound resolves five and Chebyshev resolves none. These secondary outcomes do not change the scalar result on the primary panel.

The scalar score radii are dominated by approximation, not rounding: the exact-spectrum residual terms range from .093717 to .095246 for Chebyshev and .022146 to .047756 for Lloyd, while their noise terms are at most 4.518×10^{-6} and 7.513×10^{-6}.

The primary normalized matrix ranks are 64 with one log and 192 with three. The remaining target-operator norm $\|T_L(I - A_n^\dagger A_n)\|_{\mathrm{op}}$, using the same truncated pseudoinverse, ranges from .046393 to .062878 with one log and from 6.5027×10^{-6} to 8.9819×10^{-6} with three; this one-log operator does not exactly identify the linear target. Direct-GD and spectral targets agree within 9.5162×10^{-16} in Frobenius norm. All 2,976 candidate coverage checks and 96 distinct prefix/encoding checks

Table 11: Primary-panel score scales, all multiplied by 10^3; the tolerance therefore equals 10 in this table. U_B is the common baseline upper endpoint, $U_{C,p}$ is the Chebyshev product candidate's own upper endpoint, $U_{S,s}$ is the scalar spectrum-suite upper endpoint, and I_M is the matrix-suite interval. Baseline and scalar-suite lower endpoints are zero. Displayed decimals are shortened; decisions use full precision.

Root	s_L	U_B	$U_{C,p}$	$U_{S,s}$	I_M
1	0.7308	140.543	160.351	42.144	$[0.6887, 0.7731]$
2	0.5771	131.329	159.017	27.766	$[0.5444, 0.6099]$
3	0.7801	145.488	160.677	45.970	$[0.7400, 0.8204]$
4	0.5221	134.614	159.172	22.754	$[0.4889, 0.5553]$
5	0.6939	145.335	158.275	23.973	$[0.6581, 0.7296]$
6	0.6926	141.146	158.571	42.758	$[0.6598, 0.7255]$
7	0.7652	131.989	159.287	44.879	$[0.7308, 0.7993]$
8	0.8054	140.897	159.766	48.555	$[0.7681, 0.8428]$

pass, with zero erroneous definite decisions. These are float64 analytic-bound evaluations, not outward-rounded proofs. All true score changes are below .01 (maximum .0012415): there are no fail-decision examples, and longer-horizon scalar abstention remains. Computation took 49.20 seconds with one CPU thread; setup and transfer are excluded.

L.4 Matched scalar-block and matrix recovery

We use a secondary reanalysis of the eight previously evaluated pools to test the recovery coefficient family with the other comparison choices held fixed. The authors had already seen the original targets and outcomes; this is not new blind data. The original scalar interpolation, matrix comparison, and restricted-radius results remain unchanged.

Common information and constraints. Both families receive exactly the same $H_k, p_k, W_0, \delta_{\mathrm{fit}}$, acquired logs $\widetilde{y}_m$, allowances ε_m, and C_{probe}. Neither receives labels, raw features, B_k, child endpoints, or target-horizon aggregates. The observation blocks A_m, z_m and target $D_L = d_L - T_L V$ are those defined above, with $C_k = H_k - \gamma I$. Both use the same relaxation

$$B_k = C_k^{1/2} V_k, \quad \|V_k\|_F \le 1, \qquad \|A_m V - z_m\|_F \le \varepsilon_m \quad (0 \le m \le q).$$

Here V stacks the V_k, $q \in \{0, 1, 3\}$, and even $q = 0$ includes the fitted-head block $A_0 = [p_k C_k^{1/2}]_k$, $z_0 = \bar{H} W_0$, $\varepsilon_0 = \delta_{\mathrm{fit}}$; the head is not assumed exactly stationary. The same rounding and fit allowances specified above apply.

Identical objective, different coefficient families. Normalize each *whole* observation block by one scalar: $s_m = \|A_m\|_F / \sqrt{d}$ (set $s_m = 1$ for a zero block), $A'_m = A_m / s_m$, and $A' = \mathrm{vstack}_{m=0}^{q} A'_m$. Unlike coordinatewise scaling, this preserves scalar-block membership. For each q and λ, minimize the same objective over either family:

$$J_\lambda(Q) = \left\| T_L - \sum_{m=0}^{q} Q_m A'_m \right\|_F^2 + \lambda \sum_{m=0}^{q} \|Q_m\|_F^2, \tag{67}$$

$$\mathcal{F}_{\mathrm{scalar}} = \{Q : Q_m = a_m I_d,\ a_m \in \mathbb{R}\},$$

$$\mathcal{F}_{\mathrm{matrix}} = \{Q : Q_m \in \mathbb{R}^{d \times d}\}.$$

The scalar problem is solved by direct SVD of $[\mathrm{vec}(A'_0), \ldots, \mathrm{vec}(A'_q)]$; its penalty is $\lambda d \|a\|_2^2$ because $\|a_m I_d\|_F^2 = d a_m^2$. The matrix problem uses SVD ridge recovery. Both use the same scale $s_{\max}(A')^2$ and fixed grid

$$\lambda / s_{\max}(A')^2 \in \{0, 10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\},$$

plus zero recovery. At $\lambda = 0$, each design uses a relative singular-value cutoff 10^{-12}; positive ridge uses every singular value. The matrix family contains the scalar family, so exact positive-ridge minimization cannot increase J_λ. Separate truncations at $\lambda = 0$ do not give this guarantee, and objective ordering does not imply ordering of the operator-norm radius below. All candidates are retained, without target-based regularization or radius tuning.

Common error accounting and decisions. Map back to original units with $R_m = Q_m/s_m$. With A, z stacked in observation order, both families use

$$\widehat{D}_L = d_L - Rz, \qquad \mathcal{R}(R) = \sum_k \|(T_L - RA)_k\|_{\mathrm{op}} + \sum_{m=0}^{q} \|R_m\|_{\mathrm{op}}\varepsilon_m.$$

The first blocks are indexed by child; the second by observation, including the fit block. Each candidate yields the same score interval $[\max(0, \widehat{s} - r), \widehat{s} + r]$, with $\widehat{s} = \|C_{\mathrm{probe}}^{1/2}\widehat{D}_L\|_F$ and $r = \sqrt{\|C_{\mathrm{probe}}\|_{\mathrm{op}}}\mathcal{R}(R)$. For each family separately, cumulative suites intersect its candidates with budgets $q' \le q$, while the direct three-log suite uses $q' = 3$ alone. No matrix candidate enters the scalar suite. Intersections use the same outward margin $10^{-10}(1 + \max(|\ell|, |u|, \tau))$ at each end, and the same pass/abstain/fail rule and evaluation numerical bands as above. Inconsistent intersections or uncovered targets are failures, not excluded cases.

Primary result and full grid. For $\alpha = .75$, $L = 6$, coordinate rounding 10^{-8}, and $\tau = .01$, neither family resolves any of the eight pools with zero or one log. With three logs, scalar-block recovery abstains on all eight and matrix recovery gives eight correct passes, both directly and cumulatively. The true scores range from .000522 to .000805, below the threshold. Cumulative scalar upper endpoints range from .136172 to .152247; matrix upper endpoints range from .000555270 to .000842751. Displayed values are rounded; decisions use full precision. Direct three-log matrix upper endpoints equal the cumulative endpoints on all eight pools. Although their smallest uppers use $\lambda = 0$, each positive ratio $10^{-12}, 10^{-10}, 10^{-8}$ also gives a pass by itself on every primary pool.

Table 12: Complete matched-family grid, shown as pass/abstain counts; there are no fail decisions. Counts cover two encodings, eight reused pools, two step sizes, and three horizons, not 96 independent datasets. The zero- and one-log counts agree, as do direct and cumulative counts.

τ	Family	$q = 0$ or 1	$q \le 3$	Direct $q = 3$
.01	Scalar-block	0/96	0/96	0/96
	Matrix	0/96	96/0	96/0
.05	Scalar-block	32/64	32/64	32/64
	Matrix	32/64	96/0	96/0

Table 12 retains both encodings and every $\alpha \in \{.25, .75\}$, $L \in \{4, 6, 10\}$ setting. All true scores are below .01 (maximum .001241443). At $\tau = .05$, the 32 earlier passes occur at $\alpha = .25$, $L = 4, 6$ in both encodings and are not new three-log decisions. All 5,184 candidate coverage checks pass, with zero erroneous definite decisions. These are float64 analytic-bound evaluations, not outward-rounded proofs. The comparison supports a benefit of matrix coefficients under this fixed Frobenius surrogate and candidate grid, not scalar information impossibility or optimal recovery. It neither establishes utility of the original scalar bound nor tests near-threshold or true-fail cases.

M From endpoint uncertainty to probe-level behavior

An RMS score tolerance does not specify which classifications change. We therefore apply the unchanged ridge endpoint forecasts to a secondary, label-free question: does the partition update preserve at least 99% of the union update's classifications on the fixed probe? The readout was specified before computing its per-example margins and outcomes. It was not part of the original ridge experiment's primary specification.

M.1 Available information and behavior bounds

In addition to the public covariance matrices, fitted head, fit-error bound and saved forecasts, this readout supplies the individual unlabeled probe features x_i. This is more information than their aggregate covariance. Training labels and child cross-moments are unavailable to the predictor; the evaluator accesses the original cross-moments only after all probe predictions have been saved. The interface describes which arrays the computation uses, not a privacy mechanism.

Write $\|\bar{H}W_0 - \bar{B}\|_F \le \delta_{\mathrm{fit}}$. For horizon L, the union displacement is

$$U_L - W_0 = \phi_L(\bar{H})(\bar{H}W_0 - \bar{B}).$$

Hence its class-score vector at x_i lies in a ball centered at $r_i = x_i^\top W_0$ of radius $b_i = \delta_{\mathrm{fit}}\|\phi_L(\bar{H})x_i\|_2$. An endpoint-gap forecast $\widehat{D}_j$ with $\|D_L - \widehat{D}_j\|_F \le R_j$ similarly gives a partition-score ball centered at $c_{ij} = x_i^\top(W_0 + \widehat{D}_j)$, with radius $e_{ij} = b_i + R_j\|x_i\|_2$. The fitted head is not silently substituted for the union endpoint.

For a score center c and radius e, define

$$\mathcal{S}(c, e) = \{a : \max_k c_k - c_a \le \sqrt{2}e\}.$$

If $\|z - c\|_2 \le e$, any maximizing class of z belongs to this set: the difference of two score coordinates changes by at most $\sqrt{2}e$. Let $\mathcal{U}_i = \mathcal{S}(r_i, b_i)$ and $\mathcal{P}_i = \bigcap_j \mathcal{S}(c_{ij}, e_{ij})$. When the endpoint bounds hold, these sets contain the respective true classifications. Consequently the actual disagreement fraction q satisfies

$$\frac{1}{N}\sum_i \mathbf{1}\{\mathcal{P}_i \cap \mathcal{U}_i = \varnothing\} \le q \le 1 - \frac{1}{N}\sum_i \mathbf{1}\{\mathcal{P}_i = \mathcal{U}_i = \{a\} \text{ for some } a\}. \tag{68}$$

These are simultaneous finite-probe bounds; no independence assumption on examples is needed. For any labels on the same probe, $|\operatorname{Acc}(P_L) - \operatorname{Acc}(U_L)| \le q$. A small upper bound therefore also limits the absolute accuracy change, without observing labels. A positive disagreement lower bound, however, does not establish an accuracy decrease.

The interval establishes preservation when its upper endpoint is at most $\rho = .01$, excludes that contract when its lower endpoint exceeds ρ, and otherwise abstains. We use the same rule at every root.

M.2 Fixed readout and numerical evaluation

The readout retains all eight ridge roots, the original 10,000 probe feature rows and order, the same 512-to-64 projection and common scale, $\alpha = .75$, and coordinate rounding at 10^{-8}. Horizon $L = 6$ is primary; $L = 4, 10$ are secondary. Probe covariance is checked against each original public covariance before prediction. No encoder or head is trained and no test labels are read.

The zero/one-log baseline intersects all 20 original valid scalar and matrix candidate balls. The richer interface adds all nine three-log matrix candidates, rather than selecting a favorable center after seeing classifications. Original Frobenius radii, not score-norm radii, are used in the transfer above. All 24 root–horizon cases remain in the readout.

For float64 computations we add $10^{-10}(1 + \|c\|_2)$ to each logit-ball radius before forming possible-label sets. Ties remain possible; actual argmaxes select the smallest class index. Empty intersections or failed coverage stop the readout and remain failures. Numerical evaluation checks each candidate ball, every row's possible-label membership, both fraction bounds, and every definite decision. Union displacements and partition gaps are independently reconstructed from the original quadratic sufficient statistics and compared with the saved direct and spectral endpoints. Saved-only verification used the archived probe, forecasts, endpoints and row-level outcomes. This is a numerical audit of real-arithmetic bounds, not an outward-rounded proof.

New decisions count only when the zero/one-log baseline abstains and the richer interface is correct and definite. Acceptances and rejections are reported separately. The primary usefulness criterion requires at least four of eight such changes at $L = 6$, with no integrity or coverage failure. A separate fixed selection chooses the largest accepted horizon in $\{4, 6, 10\}$, falling back to $L = 1$ by exact one-step mass invariance. It assumes no monotonicity in the horizon and implies neither measured runtime savings nor improved task accuracy.

M.3 Results and horizon choice

At the primary horizon $L = 6$, the three-log interface makes eight new correct preservation decisions, versus none for the zero/one-log baseline. Its disagreement upper bounds range from .0025 to

.0069; actual disagreement ranges from .0012 to .0044. Thus every upper bound is below the fixed .01 tolerance. All 24 cases pass the candidate-ball, row-membership, aggregate-interval and definite-decision checks.

At $L = 4$ the richer interface also accepts every root. At $L = 10$, all its intervals remain unresolved: their upper endpoints range from .024 to .035 although actual disagreement is .0025–.0068. This is a limit of the readout, not an observed failure of classification preservation. No setting produces a correct rejection. Table 13 reports all three horizons; Table 14 retains every root and both interfaces.

Table 13: Classification-preservation readout on eight roots at each horizon. A/R/? denote acceptance, rejection and abstention at disagreement tolerance .01. All numbers in the last two columns are fractions of the same 10,000 probes, not percentages. The median upper endpoint is for the three-log interface. Horizon six is the primary comparison.

L	Baseline A/R/?	Three-log A/R/?	New A	New R	Median q	Median upper
4	0/0/8	8/0/0	8	0	0.00210	0.00235
6	0/0/8	8/0/0	8	0	0.00295	0.00545
10	0/0/8	0/0/8	0	0	0.00490	0.03110

Selecting the largest accepted tested horizon chooses six for all roots; the baseline falls back to the one-step identity on all of them (Table 15). This is a decision under the specified probe-level constraint, not a claim that six is optimal or that the rejected-by-abstention horizon ten violates it. Since classification agreement bounds absolute empirical accuracy change for any labels, the accepted choice also limits that change to one percentage point on this probe. No test labels or accuracy measurements are used to make it.

Table 14: Complete classification-preservation intervals. Root numbers are one-based; all bounds and observed disagreements are fractions. A/R/? denote acceptance, rejection and abstention. "New" requires baseline abstention and a correct definite richer decision.

Root	L	Actual q	Base lower	Base upper	Base	Rich lower	Rich upper	Rich	New
1	4	0.00210	0.00000	0.99470	?	0.00190	0.00230	A	yes
2	4	0.00070	0.00000	0.99910	?	0.00060	0.00070	A	yes
3	4	0.00210	0.00000	0.99390	?	0.00170	0.00240	A	yes
4	4	0.00210	0.00000	0.99240	?	0.00190	0.00220	A	yes
5	4	0.00270	0.00000	0.99920	?	0.00240	0.00280	A	yes
6	4	0.00160	0.00000	0.99510	?	0.00130	0.00190	A	yes
7	4	0.00320	0.00000	0.98940	?	0.00290	0.00320	A	yes
8	4	0.00220	0.00000	0.99690	?	0.00210	0.00240	A	yes
1	6	0.00310	0.00000	1.00000	?	0.00170	0.00560	A	yes
2	6	0.00120	0.00000	1.00000	?	0.00040	0.00250	A	yes
3	6	0.00290	0.00000	0.99960	?	0.00110	0.00560	A	yes
4	6	0.00300	0.00000	1.00000	?	0.00140	0.00530	A	yes
5	6	0.00440	0.00000	1.00000	?	0.00270	0.00690	A	yes
6	6	0.00200	0.00000	1.00000	?	0.00090	0.00390	A	yes
7	6	0.00440	0.00000	0.99990	?	0.00200	0.00620	A	yes
8	6	0.00280	0.00000	1.00000	?	0.00170	0.00410	A	yes
1	10	0.00480	0.00000	1.00000	?	0.00000	0.03400	?	no
2	10	0.00250	0.00000	1.00000	?	0.00000	0.02410	?	no
3	10	0.00500	0.00000	1.00000	?	0.00010	0.03500	?	no
4	10	0.00510	0.00000	1.00000	?	0.00000	0.03130	?	no
5	10	0.00680	0.00000	1.00000	?	0.00010	0.03300	?	no
6	10	0.00430	0.00000	1.00000	?	0.00000	0.02680	?	no
7	10	0.00570	0.00000	1.00000	?	0.00020	0.03090	?	no
8	10	0.00370	0.00000	1.00000	?	0.00020	0.02400	?	no

Execution and reproduction. The completed readout took 65.43 host-CPU seconds on the A40 server, including saved-result verification, without GPU use or further training. An initial attempt

Table 15: Largest accepted tested horizon under the fixed 99% agreement condition. The baseline's horizon one follows analytically from mass invariance; it is not an additional measured endpoint.

Root	Baseline horizon	Three-log horizon	Actual disagreement
1	1	6	0.00310
2	1	6	0.00120
3	1	6	0.00290
4	1	6	0.00300
5	1	6	0.00440
6	1	6	0.00200
7	1	6	0.00440
8	1	6	0.00280

timed out before prediction completed and before private evaluation. Repeated process-tree monitoring was the cause; sampling that monitor every .1 seconds, while retaining every-call deadline checks and the same 180-second cap, enabled the separately recorded completion. The original failure is retained alongside the execution-only revision. The scientific arrays, candidates, formulas and thresholds were unchanged; both execution records are retained locally.

N Cold-start detection of a behavior change

The fitted-head readout establishes preservation but contains no actual violations. To test the other decision outcome, we specify one cold-start readout on the same eight ridge loss pools. The input is zero; all data, child assignments, masses, feature transformations and 10,000 unlabeled probes remain unchanged. We retain $\alpha = .75$, $L = 6$, coordinate rounding 10^{-8} and disagreement tolerance $.01$. This is a secondary algebraic replay, specified before its transformed target outcomes were computed; it is not an independent dataset or a new training replication.

N.1 A cold input changes the available information

Write the ridge child objective as $F_k(W) = \frac{1}{2}\operatorname{tr}(W^\top H_k W) - \operatorname{tr}(B_k^\top W) + \text{const}$, with $C_k = H_k - \gamma I$ and $S_k = C_k^{1/2}$. For the original one-hot targets, $B_k = S_k V_k$ for some $\|V_k\|_F \le 1$. The covariance matrices and rounded early aggregate endpoints are supplied to the predictor; B_k, labels, the original fitted head and its stationarity bound are not. Replacing the fitted input by zero while retaining its stationarity equality would describe a different problem.

Use the positive update polynomial $F_m^+(H) = \eta \sum_{t=0}^{m-1}(I - \eta H)^t = -\phi_m(H)$, and define

$$\mathsf{A}_m = [p_k F_m^+(H_k) S_k]_k, \qquad J = F_L^+(\bar{H})/\eta, \qquad \mathsf{T} = \mathsf{A}_L - J\mathsf{A}_1.$$

Stacking the V_k gives $y_m = \mathsf{A}_m V$, $U_L = J y_1$ and $D_L = \mathsf{T}V$. For q supplied logs, let $\mathsf{A} = [\mathsf{A}_1^\top, \dots, \mathsf{A}_q^\top]^\top$. Given a linear gap recovery R, its partition recovery is $R_P = R + JE_1$, where E_1 extracts the first log. Thus

$$\widehat{P}_L = R_P \widetilde{y}, \tag{69}$$

$$\|P_L - \widehat{P}_L\|_F \le \sum_k \|(\mathsf{T} - R\mathsf{A})_k\|_{\mathrm{op}} + \sum_{m=1}^{q} \|(R_P)_m\|_{\mathrm{op}} \varepsilon_m. \tag{70}$$

The subscripts index child and log blocks, respectively. The shared first-log error is combined inside R_P before taking norms, for both the baseline and richer interface. The union score at probe x_i has center $x_i^\top J \widetilde{y}_1$ and radius $\varepsilon_1 \|x_i^\top J\|_2$. Applying the possible-class construction of Appendix M yields the same disagreement interval.

The one-log baseline intersects twelve candidates: a zero-partition prior, all nine matrix-one candidates (zero, pseudoinverse and seven fixed regularizations), and both scalar-one node rules used in the earlier comparison. The richer interface adds all nine matrix-three candidates. The matrix formulas, row scaling, relative singular-value cutoff and regularization ratios are unchanged; only the observation equations and joint-noise accounting reflect the cold input. The comparison concerns these declared implementations, not the best possible one-log estimator.

N.2 TRANSFORMATION, EVALUATION, AND OUTCOME

Let $W_\star$ denote the original fitted input and put $M_m = \sum_k p_k (I - \eta H_k)^m$, $N_m = (I - \eta \bar{H})^m$. The original early logs are displacements from $W_\star$. Their cold counterparts are

$$y_m^{\mathrm{cold}} = y_m^{\mathrm{warm}} + (I - M_m) W_\star.$$

We transform the original unrounded logs, then round once. Each Frobenius error budget includes its original log allowance, the fresh rounding bound $\sqrt{64 \cdot 100}\, 10^{-8}/2$, and the fixed numerical allowance $10^{-10}(1 + \|y_m^{\mathrm{warm}}\|_F + \|W_\star\|_F + \|M_m W_\star\|_F)$. Numerical center/radius and score allowances are specified in the supplied protocol; they are not outward-rounded machine proofs.

All eight forecasts are saved before evaluation reads the original target endpoints. The evaluator then computes

$$U_L^{\mathrm{cold}} = U_L^{\mathrm{warm}} - N_L W_\star, \qquad P_L^{\mathrm{cold}} = P_L^{\mathrm{warm}} - M_L W_\star.$$

Their difference is checked against both saved direct and spectral gaps after adding $(N_L - M_L)W_\star$. In particular, we do not replace U_L^{warm} by $W_\star$: the original nonzero stationarity residual is retained. This transformation reuses the original direct/spectral validation rather than claiming a new direct optimization run.

All eight cold cases violate the 99% agreement condition. Three logs identify seven violations that the one-log baseline leaves unresolved; one remains unresolved even though it also violates the condition. There are no incorrect definite decisions or numerical coverage failures. Table 16 gives every case, including the unresolved one. The fixed success criterion required at least four additional correct decisions, including one rejection, with no wrong decision or coverage failure. It is satisfied without changing the input, horizon, threshold, rounding rule or candidate family after observing outcomes.

Table 16: Cold-start replay on the same eight ridge loss pools. Fractions, not percentages, of disagreement on 10,000 probes; threshold .01. The one-log interval is $[0, 1]$ and abstains in every row.

Root	Actual disagreement	Three-log lower	Three-log upper	Decision
0	.0240	.0105	.0426	Reject
1	.0259	.0117	.0424	Reject
2	.0228	.0103	.0457	Reject
3	.0148	.0049	.0354	Abstain
4	.0304	.0138	.0548	Reject
5	.0275	.0134	.0466	Reject
6	.0422	.0286	.0584	Reject
7	.0240	.0130	.0372	Reject

Preparation, prediction, evaluation and saved-only verification took 14.36 host-CPU seconds on the A40 server, with no GPU training. The warm and cold readouts provide examples of both decision outcomes on these loss pools; neither a rejection nor the observed disagreement establishes a decrease in accuracy.

O PROSPECTIVE PREDICTION-BOUNDARY EVALUATION

O.1 PROTOCOL FIXED BEFORE ENDPOINT EVALUATION

We use the same normalized FEMNIST digit-parity logistic objective as Appendix J.5. From 3,374 writers with at least 32 examples, we exclude the 298 writers found in the documented earlier smooth prediction and coverage panels, including their numerical-check and failed attempts. A fixed salted ordering selects eight numerical-check writers and 32 evaluation writers from the remaining 3,076. This establishes disjointness from those smooth panels, not a claim that the writers have never appeared in any historical task-model training.

Each writer has four children per candidate, with at least four examples per child. Eight hash candidates use deterministic sample-level hashes; eight label-block candidates sort by parity label with a deterministic within-label hash, then split into four contiguous blocks. No candidates or

writers are replaced based on derivatives or endpoints. Reference inputs are obtained after 25 and 100 union-GD steps at $0.25/L_{\mathrm{union}}$. For each input, the same fixed grid uses $E \in \{2, 5, 10, 20\}$ and $\alpha = \eta L_* \in \{.1, .2, .4, .8\}$, with $L_* = \max(L_{\mathrm{union}}, \{L_k\})$ over all candidate children. The 16,384 evaluation maps and 4,096 numerical-check maps complete this grid.

We designate $\|A\|$ as the prediction score before computing these endpoints. Comparators are V_g, the frozen-quadratic gap norm, and a fixed random score. For each candidate, input derivatives are saved before its endpoint evaluation. The frozen-quadratic maps use the exact input gradient and Hessian through Hessian–vector products, with the same masses, input, step size and horizon. Both nonlinear and quadratic endpoints use prefix reuse without altering any update. We retain D, D_{quad} and Q, with $R_{\mathrm{total}} = D - Q$, $R_{\mathrm{comp}} = D_{\mathrm{quad}} - Q$ and $R_{\mathrm{nonlin}} = D - D_{\mathrm{quad}}$ reproducible from those vectors.

All computations are float64 and use one CPU thread on an A40 server, with no GPU computation. Each nonlinear local trajectory is checked for finite, non-increasing loss, allowing a rounding slack of $64\epsilon_{\mathrm{mach}}(1 + |F_{\mathrm{previous}}|)$. The two input identities for A agree within the fixed tolerance. The maximum one-step gap is 5.986×10^{-15}. Independent extended-precision calculations cover every grid endpoint for all eight numerical-check writers and the first four evaluation writers. The maximum saved residual-decomposition identity error is 1.060×10^{-14}. The fixed run completed without a numerical or resource stop: 147.07 seconds of driver time and 370.09 seconds including preparation and collection. These are whole-run costs, not a cost advantage for computing A.

O.2 Estimands and fixed interpretation rules

The writer is the resampling unit. Within a writer, family, input and grid setting, all scores use the same membership-distinct candidates with resolved endpoint gaps. Resolution requires $\|D\| > 100[64\epsilon_{\mathrm{mach}}(1 + \|w\|)]$. Rank interpretation requires at least four resolved candidates with distinct measured gaps, and at least 24 eligible writers per cell. All 32 writers and all eight candidates are eligible in every evaluated cell: no duplicate, unresolved, or constant-score case is filtered out. Ordinary Spearman calculations agree with the implemented numerical-tie rule on this panel.

Correlations are averaged over writers. Percentile 95% intervals resample writers 10,000 times, with seed 20260912 plus the zero-based lexicographic $(\mathrm{family}, t, E, \alpha)$ cell ordinal. This is separate from the earlier panel's resampling in the earlier reused-writer panel. Vector-error ratios, cosines and path diagnostics first take a median over a writer's candidates and then a median over writers. The numerical supplement retains every grid cell at both reference inputs. Per-writer arrays, regret, pairwise agreement and path-curvature records remain in the local archive.

Before evaluation, a supported prediction region was defined by mean correlation $\geq .7$, its interval lower endpoint $> .5$, median relative vector error $\leq .25$, and median cosine $\geq .9$, with adequate eligible writers. Mean correlation below .7 was designated ordering degradation. The fixed stress setting was $t = 100, E = 20, \alpha = .8$; a median total-residual ratio of at least .5 marked substantial departure from the leading prediction. These are reporting conventions, not statistical tests of a universal boundary. All grid cells are retained regardless of their status, and no larger step size, horizon or replacement cohort was appended.

The completed grid has 28 supported cells, 30 with correlation at least .7 but relative vector error above .25, and six ordering-degradation cells, all six in label-block partitions. In both input states these are $(E, \alpha) = (10, .8), (20, .4), (20, .8)$. The stress setting reaches a median total-residual ratio of 2.475 for hash and 4.281 for label-block. Median composition-residual ratios are 2.425 and 3.979; median nonlinear-residual ratios are .198 and .349. The corresponding median child-path curvature changes are 23.1% and 81.4%, measured as relative changes in the per-example logistic weights $p(1 - p)$, not Hessian operator-norm changes. The vector decomposition distinguishes these sources; their norms need not add. Rank accuracy and relative vector accuracy also need not degrade together, as the hash stress comparison demonstrates.

The complete 25-step and 100-step tables are provided as tabular data, including every stress setting and unsuccessful comparison.

P MARGINAL-CURVATURE COMPARISON ON THE PREDICTION GRID

This comparison adds $C_{\mathcal{P}} = \sqrt{V_g V_H}$ and V_H to the saved inputs and endpoints of Appendix O. The addition was designed after that endpoint study, with its calculation and reporting protocol fixed before computing the new scores. It uses the same writers, checkpoints, child memberships, masses, candidate mask, and full horizon–step-size grid. No new training or endpoint trajectories are involved. The eight development writers check the numerical calculation; the 32 evaluation writers supply the comparisons.

Computing the centered-Hessian norms. For a writer with N examples and feature matrix $X \in \mathbb{R}^{N\times 785}$, let $X^\top = QR$ be an untruncated thin QR factorization with all N columns. At the saved input, let $q_i = \sigma(x_i^\top w)(1-\sigma(x_i^\top w))$. For child j with n_j examples, define

$$B_j = R\,\mathrm{diag}\left(q_i\left(\frac{\mathbf{1}\{i\in j\}}{n_j} - \frac{1}{N}\right)\right)R^\top.$$

Then $H_j - \bar{H} = QB_jQ^\top$, and

$$\|H_j - \bar{H}\|_{\mathrm{op}} = \max_{\lambda\in\mathrm{spec}(B_j)}|\lambda|, \qquad V_H = \sum_j \frac{n_j}{N}\|H_j - \bar{H}\|_{\mathrm{op}}^2.$$

The remaining ambient eigenvalues are zero. Because B_j can be indefinite, both spectral extremes enter the norm. This is a change of basis, not a truncated-spectrum or randomized estimate. Writer sample sizes range from 43 to 117, allowing the complete reduced eigendecompositions to be computed directly.

The calculation checks the saved gradients, pairing vectors, masses, QR reconstruction and orthogonality, eigenpair residuals, and the inequality $\|A\| \le C_{\mathcal{P}}$. It also compares reduced and full 785×785 norms for the predetermined 128 development-child cases (both inputs, both families, and candidate zero for all eight writers). All checks pass. Reduced matrices, spectra, extreme eigenvectors, and gradient coordinates are retained for all 1,280 inputs. The computation takes 52.15 seconds on one CPU thread of the A40 server, with no GPU work. This timing covers the added marginal-curvature calculation, not the construction of a certified remainder bound.

Comparison and uncertainty. All 64 evaluation settings retain the original 32 writers and eight candidates per writer. We correlate each score with the observed endpoint gap norm within writer and family, then average over writers. The paired contrast subtracts each writer's $C_{\mathcal{P}}$ correlation from its $\|A\|$ correlation before averaging. Intervals use 10,000 writer-bootstrap draws, seed 20260912 plus the zero-based lexicographic index of $(\text{family}, t, E, \alpha)$. They describe this subsequent comparison pointwise and are not multiplicity-adjusted claims of superiority.

The mean paired difference is positive in every setting, ranging from .0573 to .5826; all pointwise intervals exclude zero. No score is undefined and no writer is removed. These results show that the observed benefit of pairing on this grid is not explained by adding marginal curvature dispersion to gradient dispersion. They do not turn ordering into a consequence of the smooth bound: at the fixed label-block stress point, the correlations of both $\|A\|$ and $C_{\mathcal{P}}$ are low (.4494 and .1674). The numerical supplement contains the complete grid; writer-level arrays and development records remain in the local archive. Aggregate improvements coexist with 107 negative and 38 zero differences among the 2,048 evaluation writer–setting pairs. The development comparisons also include a reversal: for hash partitions at $t = 25$, $E = 20$, and $\alpha = .8$, the mean difference is $-.0833$ with interval $[-.1815, -.0030]$; three other development intervals cross zero.

The complete grid is supplied as tabular data rather than repeated here.

Q CANDIDATE DIRECTION, MAGNITUDE, AND ORDERING

High ordering correlation need not imply an accurate vector prediction. To examine whether the logistic hash result is explained by nearly collinear candidate gaps, we analyze the saved signed vectors for all 32 formal writers and all 64 settings. This is a post-hoc descriptive analysis; no endpoints are recomputed or candidates reselected.

Table 17: Pairing versus marginal summaries at the two fixed comparisons ($t = 100$). Scores are mean writer-wise Spearman correlations with the endpoint gap norm. The last column is a difference of correlations, not a difference of score magnitudes; brackets give writer-paired, pointwise 95% bootstrap intervals. The $C_{\mathcal{P}}$ and V_H comparisons were added after the endpoint study.

Family	(E, α)	$\|A\|$	$C_{\mathcal{P}}$	V_g	V_H	$\|A\| - C_{\mathcal{P}}$ ranking difference
Hash	$(2, 0.1)$	0.9993	0.8408	0.8475	0.2455	0.1585 [0.1213, 0.1972]
Label-block	$(2, 0.1)$	0.9940	0.4115	0.0342	0.4182	0.5826 [0.4926, 0.6749]
Hash	$(20, 0.8)$	0.8936	0.8363	0.8542	0.2522	0.0573 [0.0223, 0.0923]
Label-block	$(20, 0.8)$	0.4494	0.1674	-0.0841	0.1897	0.2820 [0.1689, 0.3884]

For each writer and setting, let $\mathsf{D} \in \mathbb{R}^{8\times d}$ stack the candidate gaps and let D_c subtract their row mean. We compute signed cosines between candidate pairs, the energy fraction $\sigma_1(\mathsf{D})^2/\|\mathsf{D}\|_F^2$, and its centered counterpart. The uncentered fraction measures a dominant common direction; the centered fraction instead measures whether candidate differences are mostly one-dimensional. Neither quantity alone determines norm ordering. We also compare D and Q within each candidate, separating directional alignment from the norm ratio $\|Q\|/\|D\|$.

Table 18 reports medians over writers after within-writer aggregation. Cosines use the original numerical-resolution threshold and retain signed, not absolute, values. The local archive contains the complete 64-setting readout and all 2,048 writer groups. Independent direct SVD and Gram-matrix calculations agree, and correlations recomputed from vectors reproduce the original summary.

Table 18: Geometry of logistic candidate gaps at the two fixed comparisons, with reference step $t = 100$. Pair cosine is the within-writer median over candidate pairs. Raw/centered PC1 are squared singular-value energy fractions. Other geometry columns first take candidate medians within a writer. All geometry entries then take the median over 32 writers; ρ remains the mean writer-wise Spearman correlation.

Family	(E, α)	ρ	Pair cosine	Raw PC1	Centered PC1	D/Q cosine	$\|Q\|/\|D\|$
Hash	$(2, .1)$	.9993	.2901	.4589	.4265	1.0000	1.0027
Hash	$(20, .8)$	.8936	.4670	.5405	.3510	.9173	3.3344
Label-block	$(2, .1)$	.9940	.9811	.9831	.4237	1.0000	1.0141
Label-block	$(20, .8)$	.4494	.9671	.9732	.3510	.9641	5.1965

At the stress comparison, hash gaps are not concentrated along a common ray: median pair cosine is .4670, and the first centered component accounts for 35.10% of candidate-variation energy. Within each candidate, D and Q retain substantially stronger directional agreement (.9173), while the median predicted-to-observed norm ratio is 3.3344. Thus retained ordering coexists with directional information and poor magnitude calibration, rather than with near-collinearity among candidates.

Label-block shows the converse warning. Its candidate gaps share a strong common direction, but their norm ordering is weaker. Only 2.70% of uncentered energy remains after centering, and the first centered component again accounts for 35.10% of that variation. A common direction therefore does not make candidate differences one-dimensional or ensure accurate ordering. These geometric observations distinguish possible explanations; they do not establish a causal reason for the correlations.

R Shared neural model and prediction definitions

The longer-trained evaluation below and the earlier near-initialization comparison use the same model, initialization, and input-local scale. Their reference states and resulting prediction behavior are reported separately.

Model and initialization. Nonoverlapping 4×4 average pooling maps each 28×28 image to 49 features, each divided by seven. Pooling is fixed preprocessing. The network has eight hidden units with exact-erf GELU activation (Hendrycks & Gimpel, 2016), one scalar output and biases in both layers: 409 trainable parameters. It minimizes mean squared loss $\frac{1}{2}(f_\theta(x) - y)^2$, where odd

digits have target $+1$ and even digits -1. There is no regularization, normalization layer, dropout or persistent optimizer state. Every writer starts from the same PCG64 initialization with seed 20260911: hidden weights uniform on $[-1/7, 1/7]$, output weights uniform on $[-1/\sqrt{8}, 1/\sqrt{8}]$, and zero biases.

An input-local scale, not a global certificate. Write $z = (x, 1)$, $u_j = q_j^\top z$, $a(u) = u\Phi(u)$, $f_\theta = \sum_j v_j a(u_j) + b$, and $r = f_\theta - y$. The sample loss satisfies

$$\nabla \ell = rJ, \qquad \nabla^2 \ell = JJ^\top + rH_f, \quad J = \nabla_\theta f_\theta,$$

so its Hessian need not be positive semidefinite. With $s^2 = \|z\|^2$, the analytic quantities

$$\|J\|^2 = s^2 \sum_j v_j^2 a'(u_j)^2 + \sum_j a(u_j)^2 + 1,$$
$$\|H_f\|_F^2 = s^4 \sum_j v_j^2 a''(u_j)^2 + 2s^2 \sum_j a'(u_j)^2$$

give $\|\nabla^2 \ell\|_{\mathrm{op}} \le \|J\|^2 + |r|\|H_f\|_F$. At each saved input, L_{in} is the maximum subset-mean bound over the union and all 64 children across the 16 candidates. All candidates share $\eta = \alpha / L_{\mathrm{in}}$. Unlike the logistic experiment's L_*, this bounds curvature at the input only; equal α is not a matched global-smoothness regime across the two models.

Matched maps and residuals. Both neural evaluations use two partition families, eight variants, $K = 4$, two reference inputs, $\alpha \in \{.1, .2, .4, .8\}$ and $E \in \{2, 5, 10, 20\}$. Every child keeps its empirical mass and uses deterministic full-batch updates. Analytic float64 gradients and Hessian–vector products compute A, Q, D, and the exact frozen-quadratic D_{quad} defined in Section 6. The quadratic Hessian is evaluated only at the input; the nonlinear arm recomputes its gradient at each step. We retain the signed identity $D - Q = (D_{\mathrm{quad}} - Q) + (D - D_{\mathrm{quad}})$, including $D_{\mathrm{quad}} = Q$ at $E = 2$.

The writer-wise estimands and fixed interpretation criteria are those in Appendix O; the neural scale above is input-local, not the logistic global curvature bound. The complete near-initialization comparison remains in the companion, including its failed quantitative stress prediction and its small measured training movement.

S Pairing prediction after longer neural training

Table 19: Prediction at the fixed 20,000-step neural states. Correlations are writer means with $\|D\|$. Relative error $\|D - Q\|/\|D\|$ and nonlinear residual $\|D - D_{\mathrm{quad}}\|/\|D\|$ are medians of writer medians.

Family	(E, α)	ρ_A	ρ_{V_g}	ρ_{quad}	Relative error	Nonlinear residual
Hash	$(2, 0.1)$	0.9799	0.7202	0.9799	0.066	0.066
Hash	$(20, 0.8)$	0.7470	0.6949	0.7210	9.758	0.899
Label-block	$(2, 0.1)$	0.9487	-0.4658	0.9487	0.092	0.092
Label-block	$(20, 0.8)$	0.4100	-0.0930	0.2961	4.025	0.584

The earlier near-initialization experiment leaves open whether its prediction behavior persists after substantial learning. This extension keeps the same 409-parameter GELU model, initialization, eight development and 32 formal writers, and all candidate memberships. It fixes a longer training horizon before computing new partition endpoints. The original results remain a separate historical comparison, not additional independent writers. Appendix R specifies the shared model, initialization, input-local scale, and residuals.

Fixed reference trajectory. Each writer takes 20,000 union full-batch GD updates at the unchanged constant rate $\eta_{\mathrm{ref}} = .25/L_{\mathrm{union}}(\theta_0)$. The analyzed states are steps 2,000 and 20,000; the latter is primary. Neither loss nor prediction accuracy selects a state or adjusts the rate. At each state, all candidate derivatives and the shared input-local scale L_{in} are computed before their endpoints.

Both families, eight variants, $E \in \{2, 5, 10, 20\}$ and $\alpha \in \{.1, .2, .4, .8\}$ are retained. The primary comparison is $(E, \alpha) = (2, .1)$, and the fixed stress comparison is $(20, .8)$. The definitions of Q, D_{quad}, R_{comp}, R_{nonlin} and the prediction criteria are unchanged. This is a trained-state extension on the existing cohort, not a new-data replication.

Measuring training and feature change separately. For each writer, the best constant output is $\bar{y}$, with loss $L_{\text{const}} = \frac{1}{2}\,\text{mean}(y - \bar{y})^2$. Define the fractional improvement over this reference as $G_t = (L_{\text{const}} - L_t)/L_{\text{const}}$ when $L_{\text{const}} > 10^{-12}$. Loss, thresholded training accuracy, gradient norm and displacement from initialization are also recorded. A minimum-norm affine least-squares fit on the same 49 features and intercept provides a descriptive in-sample reference, with SVD cutoff $\epsilon_{64} \max(n, 50) s_{\max}$.

Let U_t be the example-by-hidden preactivation matrix, $H_t = \text{GELU}(U_t)$, and C center each column over examples. The hidden-feature change and its nonlinear fraction are

$$B_t = \text{RMS}\big(C(H_t - H_0)\big), \qquad N_t = \frac{\text{RMS}\big(C[H_t - H_0 - \text{GELU}'(U_0) \odot (U_t - U_0)]\big)}{\max(B_t, 10^{-12})}.$$

RMS averages over all example-hidden entries. We retain both numerator and denominator, and the analogous output-linearization residual $\text{RMS}(f_t - f_0 - J_0(\theta_t - \theta_0))$. Its relative error divides by $\text{RMS}(f_t - f_0)$ and is defined only when this denominator exceeds 10^{-12}. These quantities distinguish fitting, representation movement and nonlinearity; none is a held-out performance measure.

Before execution we specified descriptive criteria $G_t \geq .10$ for meaningful fitting and $B_t \geq .01$, $N_t \geq .10$ for material nonlinear feature change. At least 24 of the same 32 formal writers must meet all three at step 20,000 to support the trained-state interpretation. These conventions never filter the prediction cohort, and missing or stopped writers do not count toward them. Continuous summaries accompany the counts.

Table 20: Training-state diagnostics on the fixed 32 formal writers; all continuous entries are defined for all 32 and are writer medians. "Trained" counts $G_t \geq .10$; "Joint" additionally requires $B_t \geq .01$ and $N_t \geq .10$. These counts do not select the prediction writers.

Steps	Loss	Reduction G	Hidden change B	Nonlinearity N	Trained	Joint
2000	0.4885	0.0202	0.0045	0.1836	3/32	3/32
20000	0.1904	0.6179	0.0504	0.4486	32/32	32/32

Substantial training and nonlinear feature change. At step 20,000, all 32 writers meet each condition and their conjunction; at step 2,000, the respective counts are 3 for training, 6 for nonlinear feature change and 3 jointly. The primary median training loss is 0.190356, with initial-relative loss reduction 0.619040 and training accuracy 0.905025. Median $G_t = 0.617863$, $B_t = 0.0504325$ and $N_t = 0.448638$ quantify the learned-state change. Their interquartile ranges are $[0.522695, 0.682697]$, $[0.0472414, 0.0539813]$ and $[0.438655, 0.456965]$, respectively. Median parameter displacement relative to initialization is 6.41108, and the median output-linearization relative error is 1.00052. Thus the primary inputs are no longer near initialization under the declared measures; all 32 writers remain in the prediction cohort.

These are fixed trained checkpoints, not demonstrated converged solutions. All 32 long training traces contain finite loss increases, which the protocol permits; 31 still contain increases in their final 1,000 updates. The primary median gradient norm is 0.453064. The affine reference has lower training loss on every writer at both checkpoints: its median loss is 0.134032, and the primary median paired neural-minus-affine loss is 0.0460280. Feature nonlinearity therefore does not establish superiority over the pooled affine features. These in-sample measurements on the same small cohort make no held-out or new-writer generalization claim.

Prediction succeeds locally and fails at stress. The primary small-step pairing correlations are 0.9799 and 0.9487 for HASH and LABEL-BLOCK, versus 0.7202 and -0.4658 for V_g. The pairing intervals are $[0.970982, 0.987351]$ and $[0.905506, 0.980655]$, with vector errors 0.0656872 and 0.0916812 and cosines 0.999303 and 0.998422. Both meet the quantitative criteria. The contrast

with V_g is descriptive; no paired significance test is claimed. At $E = 2$, $D_{\mathrm{quad}} = Q$ algebraically, so identical quadratic rankings are not a separate baseline victory.

At the fixed stress comparison, pairing correlations are 0.747024 and 0.409970, but vector errors are 9.75790 and 4.02461 and cosines are 0.769488 and 0.715014; neither family satisfies quantitative support. The corresponding compositional residual ratios are 9.36414 and 4.31582, and nonlinear residual ratios are 0.899476 and 0.583637. The quadratic map reduces vector error to these latter values but has lower ranking correlations, 0.720982 and 0.296131. It is not a universal repair. Residual norms are reported separately and are not additive.

All 64 settings have 32 eligible writers with defined ranks for every score and 32 cosine-eligible writers for both predictors; there are no stopped or unresolved endpoint rows. The joint prediction criteria hold in 13 settings: eight at step 2,000 and five at step 20,000. At step 2,000 both families qualify at $E = 2$ with $\alpha \in \{.1, .2, .4\}$ and at $(E, \alpha) = (5, .1)$. At step 20,000 both qualify at $E = 2$ with $\alpha \in \{.1, .2\}$, and LABEL-BLOCK additionally qualifies at $(5, .1)$. All other settings remain reported; the primary comparison is not replaced by a more favorable cell.

Integrity and full-grid reporting. The experiment uses analytic NumPy/SciPy float64 derivatives on one CPU thread of the A40 host. At both analyzed checkpoints, independent PyTorch gradients, Hessian–vector products and a fixed two-step matched map are checked on the two feasibility writers and the first four formal writers. A 25-step replay from saved anchors at steps 1,975 and 19,975 checks local trajectory consistency; it is not a second implementation of the entire 20,000-step trajectory. The same resolution, duplicate-membership and writer aggregation rules apply. Pointwise 95% intervals use 10,000 writer-bootstrap draws. All 64 formal settings, including undefined quantities and stopped paths, are retained. Correlation alone does not establish quantitative prediction support: the vector-error, direction and writer-coverage conditions also apply.

The local archive preserves parameters, scalar training traces, derivatives, diagnostics and endpoint vectors. The submission supplies the complete fixed-grid table, not these arrays or the full training implementation. Image-derived calculations require the original benchmark.

The scientific computation completed once in 438.947 seconds on the A40-host CPU, with no GPU computation. The original transfer exceeded its resource allowance; subsequent byte-verified retrieval recovered the existing complete results without scientific recomputation. This recovery does not make the original attempt resource-compliant. The local execution record distinguishes this transfer deviation from the scientific results.

The numerical supplement contains all 64 settings, not just supported ones.

T INDEPENDENT-COHORT PREDICTION AFTER MONOTONE TRAINING

This study tests whether the small-step pairing prediction extends to new writers and new initializations after substantial nonlinear training. It uses a disjoint cohort from the documented prediction studies in Appendices S and the preceding logistic evaluations. It does not change their results or pool their participants.

Data and reference training. An outcome-blind deterministic ordering of eligible FEMNIST writer IDs selects two resource-check writers and then 32 formal writers, excluding the earlier prediction cohorts. All examples of each selected writer are retained. The exact exclusion list, ordering and row memberships are supplied in the numerical supplement. This establishes separation from the documented prediction cohorts, not from every historical use of FEMNIST. The model remains $49 \to 8 \to 1$ with exact-erf GELU, 409 parameters, pooled image features, parity targets ± 1, and squared loss divided by two. Initializations use fixed seeds 20260921, 20260922 and 20260923 (seeds 1–3 in the tables), with uniform hidden weights in $[-1/7, 1/7]$, output weights in $[-1/\sqrt{8}, 1/\sqrt{8}]$ and zero biases.

Each writer and seed takes exactly 20,000 accepted union-GD updates. Starting with $.25/L_{\mathrm{union}}(\theta_0)$, the reference learning rate can only decrease, by halving until $L(\theta - \eta g) \le L(\theta) - 10^{-4}\eta\|g\|^2$. No best-checkpoint selection is used. The rate, loss, gradient norm and backtracking count are recorded throughout. Reference training is monotone by construction; this is not evidence of convergence. The in-sample fitting and feature-change descriptors G_t, B_t, N_t are those defined in Appendix S. For

each seed, all 32 writers satisfy $G_t \geq .10$, $B_t \geq .01$ and $N_t \geq .10$. These descriptors do not filter prediction cases. Table 21 reports continuous values. The affine least-squares reference has lower training loss on every writer for every seed; no model-quality advantage is claimed.

Table 21: Independent-cohort training descriptors, as writer medians. Loss reduction is relative to initialization; G_t is relative to the best constant output. All 32 writers per seed satisfy the joint fitting and nonlinear-feature-change criterion.

Seed	Loss	Loss reduction	G_t	B_t	N_t	Affine loss
1	0.1811	0.6378	0.6376	0.0423	0.5449	0.1415
2	0.1814	0.6370	0.6369	0.0422	0.5335	0.1415
3	0.1806	0.6390	0.6386	0.0428	0.5237	0.1415

Fixed-rate comparison and common scores. Each trained input is compared across $K = 4$ children, eight HASH and eight LABEL-BLOCK variants. HASH uses a fixed ordering of immutable IDs and round-robin assignment; LABEL-BLOCK sorts by binary parity target, uses ID ordering to break ties and takes contiguous blocks. Every candidate preserves the same examples and their masses. All memberships are recorded before training. The audit itself uses fixed-rate, stateless full-batch GD: $\eta = \alpha / L_{\text{in}}$ with $\alpha \in \{.1, .2, .4, .8\}$ and $E \in \{2, 5, 10, 20\}$. L_{in} is shared across union and candidate children at that input. It scales the comparison; it is not a certified global trajectory bound. All 24,576 planned finite-map comparisons completed.

Input quantities are saved before endpoint execution. They include $A_{\mathcal{P}}$, V_g, and $C_{\mathcal{P}} = \sqrt{V_g \sum_k p_k \|H_k - \bar{H}\|_{\text{op}}^2}$. For this small model, $C_{\mathcal{P}}$ uses analytic dense Hessians and symmetric extreme-eigenvalue computations. It is not a Frobenius proxy. We compare $Q = \binom{E}{2}\eta^2 A_{\mathcal{P}}$ and the frozen-quadratic gap D_{quad} with the observed vector D. At $E = 2$, $D_{\text{quad}} = Q$ is an identity, not an independent success.

Reporting unit and prediction criteria. The numerical floor is $64\epsilon_{64}(1 + \|\theta\|)$. Defects are resolved above 100 times this floor; cosine additionally requires a resolved predictor. Repeated memberships are retained and masked after their first occurrence. A writer setting needs at least four unique resolved candidates and four distinct observed levels. The same mask is used for all four scores. All 32 writers have defined rankings and directions in all 96 seed–family–setting panels. Correlations are means over writers, and vector errors and cosines are medians of writer medians. Pointwise 95% intervals use 10,000 writer-bootstrap draws within each seed; the three initializations do not create 96 independent writers.

The fixed joint criterion requires complete comparisons, at least 24 eligible writers, mean pairing Spearman at least .7 with interval lower endpoint above .5, median relative vector error at most .25, and at least 24 cosine-eligible writers with median cosine at least .9. Primary replication requires both families to satisfy this criterion and the trained-state descriptor criterion for each seed. All six primary panels at $(E, \alpha) = (2, .1)$ satisfy it. The fixed secondary $(5, .1)$ panels also satisfy it; the six $(20, .8)$ stress panels do not. Each seed supports nine of its 32 fixed settings, for 27/96 overall. No seed, writer, checkpoint or rate is replaced. At the fixed secondary comparison, D_{quad} has lower median vector error than Q in all six panels; the study does not establish superiority over frozen-quadratic prediction. The numerical supplement contains the full 96-setting table; individual writer arrays remain in the local archive. These finite-grid criteria are not multiplicity-adjusted population claims.

Table 22: Primary, fixed secondary and stress prediction panels. Intervals are pointwise writer-bootstrap intervals for ρ_A; errors are medians of writer medians. "Joint" denotes the complete ordering/error/direction criterion. The full fixed grid, including all 69 unsupported settings, is retained in the numerical supplement.

Seed	Partition	E	α	ρ_A	95% interval	Err. Q	Err. D_{quad}	Joint
1	Hash	2	0.1	0.9955	[0.9918, 0.9985]	0.0346	0.0346	Yes
1	Hash	5	0.1	0.9911	[0.9851, 0.9955]	0.1551	0.0924	Yes
1	Hash	20	0.8	0.9353	[0.9077, 0.9583]	7.3542	0.3503	No
1	Label-block	2	0.1	0.9658	[0.9360, 0.9888]	0.0423	0.0423	Yes
1	Label-block	5	0.1	0.9092	[0.8363, 0.9695]	0.1256	0.1094	Yes
1	Label-block	20	0.8	0.8326	[0.7076, 0.9189]	5.6739	0.3668	No
2	Hash	2	0.1	0.9970	[0.9940, 0.9993]	0.0379	0.0379	Yes
2	Hash	5	0.1	0.9836	[0.9747, 0.9911]	0.1761	0.0997	Yes
2	Hash	20	0.8	0.9204	[0.8921, 0.9457]	7.4799	0.3713	No
2	Label-block	2	0.1	0.9643	[0.9263, 0.9888]	0.0454	0.0454	Yes
2	Label-block	5	0.1	0.9182	[0.8556, 0.9702]	0.1208	0.1160	Yes
2	Label-block	20	0.8	0.8065	[0.6696, 0.9033]	5.9130	0.3825	No
3	Hash	2	0.1	0.9918	[0.9881, 0.9955]	0.0479	0.0479	Yes
3	Hash	5	0.1	0.9725	[0.9561, 0.9859]	0.2008	0.1265	Yes
3	Hash	20	0.8	0.8929	[0.8527, 0.9293]	7.7823	0.3927	No
3	Label-block	2	0.1	0.9643	[0.9249, 0.9896]	0.0431	0.0431	Yes
3	Label-block	5	0.1	0.8958	[0.7805, 0.9754]	0.1151	0.1108	Yes
3	Label-block	20	0.8	0.7329	[0.5804, 0.8571]	8.1165	0.3914	No

Verification and computation. The study uses one CPU thread on the A40 host, with float64 arithmetic and no GPU computation. Formal computation and summary generation took 755.1 seconds, within the fixed 2,400-second limit. The two-writer resource check precedes it and uses no prediction-success gate. Dense Hessian products are checked against analytic Hessian–vector products. Independent PyTorch checks at the final input of the first two formal writers, for all three seeds, verify gradients, Hessian–vector products, pairing and a fixed two-step comparison. A 25-update replay checks the saved final training segment, not the entire training path. One-step mass-averaging and residual identities are also checked.

The local archive contains aggregate inputs, checkpoints, training traces and endpoint vectors. Its independent checker recomputes saved-vector arithmetic, masks, writer aggregation, bootstrap intervals and training criteria; this is not raw-data training reproduction. The submission includes memberships and complete setting-level results, not the training framework. Because both cohort and training rule change from the earlier constant-rate study, their difference is not a controlled estimate of the benefit of backtracking.

U Applicability, negative controls, and computational cost

U.1 Scope of the guarantees and supporting controls

The formal certificate assumes deterministic full-batch GD, local C^3 smoothness, and a common containment region. It does not cover ReLU networks, stochastic minibatches, AdamW, persistent optimizer state, BatchNorm, privacy accountants, compression feedback, or personalization state. An interval above tolerance is unresolved unless its lower endpoint exceeds tolerance. Neither a large upper bound nor parameter separation alone proves downstream harm or mandates consolidation.

Earlier controls delimit these conclusions; their full records remain local. In the earlier reused-writer logistic evaluation, all 32 writers meet the input criterion, but the path criterion is met by only 0/32 hash and 4/32 label-block writers; the residual-stratified reanalysis is conditional, not independent confirmation. The earlier neural comparison's median writer-wise relative loss decrease is only .17561% at step 100; it is a near-initialization mechanism evaluation. The longer-trained neural evaluation retains its complete stress grid and training limitations in Appendix S.

The original ReLU remainder slope is 1.56225, outside the fixed range $[2.5, 3.5]$. GELU passes the matched activation checks while Softplus does not; retained MaxPool layers prevent interpreting

those local observations as a global smoothness certificate. The earlier $C_{\mathcal{P}}$ ranking advantage over V_g is modest and weaker within families than in aggregate. The task-model and adapter comparisons establish parameter separation with mixed downstream metric directions, not a quality guarantee; the from-scratch CIFAR-100 union reaches only 11.6% accuracy at $E = 10$ and is a failed-learning diagnostic. None of these controls establishes a general downstream benefit or extends the smooth bound to the excluded update rules.

U.2 Computational cost

The separate checkpoint-grid audit instruments the two paths directly. Over 128 roots and all twelve (E, η) settings at a checkpoint, exact endpoint simulation takes 3.09–3.14 host-CPU seconds, whereas construction of the complete B^C certificate takes 66.15–67.56 seconds in the dense exact 785-dimensional implementation. The scalar evaluations after certificate construction are negligible, but the centered-Hessian spectral norms dominate. This tested short-horizon regime therefore supports an explanatory and certifiable interface, not a runtime-acceleration claim; cached derivative information, approximate HVPs, other dimensions, and longer horizons are not evaluated by this timing comparison. Local records retain costs for the earlier controls; sums double-count concurrent execution and are not a certificate-only speed comparison.

V Evaluating the restricted noisy-observation bound

This secondary analysis evaluates Theorem 3 on the same eight ridge pools, without passing the full covariance matrices or fitted head to the predictor. It tests a sufficient bound, not the dataset-specific necessity of the observed depth. The main comparison fails to resolve any case; all results, including the limited smaller-step successes, are retained.

Information and preparation. The owner prepares a spectral set S, a gradient bound G, the step size η, early-discrepancy error budgets, and $\kappa = \sqrt{\lambda_{\max}(C_{\mathrm{probe}})}$. These scalars and the spectral set are the complete prediction interface. The owner can compute the premises from the existing covariance interface; the predictor gets no matrices, head, labels, fit bound or later endpoint. For the one-hot ridge targets, Cauchy–Schwarz gives

$$\|B_k\|_F \le \sqrt{\lambda_{\max}(H_k - \gamma I)}, \qquad G = \max_k \left\{ \|H_k W_0\|_F + \sqrt{\lambda_{\max}(H_k - \gamma I)} \right\}$$

before a stated numerical allowance. This also bounds the union gradient.

Let y_m be the recorded partition displacement and $Q_m(H) = \sum_{t=0}^{m-1} (I - \eta H)^t$. The recorded union displacement is $Q_m(\bar{H}) y_1$. Its error includes the first-log error transported by $\|Q_m(\bar{H})\|_{\mathrm{op}}$; fitted-head stationarity is not substituted for this displacement. If e_m bounds the combined record error, set

$$\varepsilon_m = \tfrac{1}{2} \left(\|y_m - Q_m(\bar{H}) y_1\|_F + e_m \right).$$

The true endpoints admit their midpoint as a common transcript within these budgets. The theorem therefore gives the RMS-score upper bound $2\kappa \mathcal{E}_r$.

Fixed grid and numerical optimization. The grid uses all eight pools, step factors $.25, .75$, targets $4, 6, 10$, and observation depths $1, 2, 3$. The existing unrounded prefixes are re-encoded with coordinate spacings $0, 10^{-8}, 10^{-6}, 10^{-4}$. These are sensitivity cases, not new independent observations. The primary spectral set is the interval spanning the child and union spectra; the secondary set is their full numerical support. The inherited primary comparison is $(\alpha, L, \tau) = (.75, 6, .01)$, spacing 10^{-8}. Tolerance $.05$ is secondary.

We solve Eq. (53) in normalized coordinate $x = \eta h$. For finite support every node is constrained. For the interval, a linear program starts from 33 cosine-spaced nodes and adds residual-polynomial extrema until the constraint violation is at most 10^{-9}, with a maximum of 64 exchanges. The reported radius recomputes the residual supremum and noise penalty, rather than reporting the solver objective alone. A feasible fixed-Chebyshev interpolant is evaluated with the same information and error budgets. Earlier feasible bounds are retained when depth increases.

The eigenvalue guard is $10^{-10}(1 + \max_k \lambda_{\max}(H_k))$. The interval is expanded accordingly; the finite-support residual adds the guard times a derivative supremum. Re-encoding includes the original log error, coordinate-rounding radius and explicit arithmetic allowances. This is numerical evaluation of an analytic bound, not outward-rounded eigenvalue enclosure or a formal machine proof. All coefficients, suprema, noise terms and solver diagnostics are saved.

Result and stopping decision. All 1,152 declared bound evaluations are available, with no observed coverage failure or false pass. At the primary comparison, depths one, two and three each resolve zero of eight cases. The three-log bounds range from .176736 to .180293; measured RMS score changes range from .000522 to .000805. The predeclared condition of at least four new primary passes therefore fails, and this branch stops without changing the thresholds or selecting another grid point as primary. Table 23 reports all spacings and spectral sets. The covariance-aware result in Table 9 uses a richer interface and is not credited to this restricted bound.

Table 23: Restricted depth audit: complete grid. Passes are counts out of eight at depths one/two/three. B_3 is the median three-log RMS-score bound. Spacings re-encode the same early logs. The primary setting is interval, $\alpha = .75$, $L = 6$, spacing 10^{-8} and $\tau = .01$.

Spacing	Spectrum	α	L	Passes (.01)	B_3	Passes (.05)
0	Interval	0.25	4	0/0/8	0.001140	0/8/8
0	Interval	0.25	6	0/0/8	0.007430	0/0/8
0	Interval	0.25	10	0/0/0	0.049583	0/0/6
0	Interval	0.75	4	0/0/0	0.026122	0/0/8
0	Interval	0.75	6	0/0/0	0.178005	0/0/0
0	Interval	0.75	10	0/0/0	0.861218	0/0/0
0	Support	0.25	4	0/8/8	0.000997	0/8/8
0	Support	0.25	6	0/0/8	0.005677	0/8/8
0	Support	0.25	10	0/0/0	0.032990	0/0/8
0	Support	0.75	4	0/0/0	0.013884	0/0/8
0	Support	0.75	6	0/0/0	0.084970	0/0/0
0	Support	0.75	10	0/0/0	0.407499	0/0/0
10^{-8}	Interval	0.25	4	0/0/8	0.001145	0/8/8
10^{-8}	Interval	0.25	6	0/0/8	0.007457	0/0/8
10^{-8}	Interval	0.25	10	0/0/0	0.049711	0/0/5
10^{-8}	Interval	0.75	4	0/0/0	0.026125	0/0/8
10^{-8}	Interval	0.75	6	0/0/0	0.178018	0/0/0
10^{-8}	Interval	0.75	10	0/0/0	0.861257	0/0/0
10^{-8}	Support	0.25	4	0/8/8	0.001002	0/8/8
10^{-8}	Support	0.25	6	0/0/8	0.005705	0/8/8
10^{-8}	Support	0.25	10	0/0/0	0.033123	0/0/8
10^{-8}	Support	0.75	4	0/0/0	0.013888	0/0/8
10^{-8}	Support	0.75	6	0/0/0	0.084984	0/0/0
10^{-8}	Support	0.75	10	0/0/0	0.407547	0/0/0
10^{-6}	Interval	0.25	4	0/0/8	0.001674	0/8/8
10^{-6}	Interval	0.25	6	0/0/2	0.010201	0/0/8
10^{-6}	Interval	0.25	10	0/0/0	0.062557	0/0/0
10^{-6}	Interval	0.75	4	0/0/0	0.026488	0/0/8
10^{-6}	Interval	0.75	6	0/0/0	0.179303	0/0/0
10^{-6}	Interval	0.75	10	0/0/0	0.865119	0/0/0
10^{-6}	Support	0.25	4	0/8/8	0.001532	0/8/8
10^{-6}	Support	0.25	6	0/0/8	0.008483	0/8/8
10^{-6}	Support	0.25	10	0/0/0	0.046474	0/0/8
10^{-6}	Support	0.75	4	0/0/0	0.014256	0/0/8
10^{-6}	Support	0.75	6	0/0/0	0.086369	0/0/0
10^{-6}	Support	0.75	10	0/0/0	0.412331	0/0/0
10^{-4}	Interval	0.25	4	0/0/0	0.020852	0/8/8

Spacing	Spectrum	α	L	Passes (.01)	B_3	Passes (.05)
10^{-4}	Interval	0.25	6	0/0/0	0.075274	0/0/0
10^{-4}	Interval	0.25	10	0/0/0	0.297782	0/0/0
10^{-4}	Interval	0.75	4	0/0/0	0.062072	0/0/0
10^{-4}	Interval	0.75	6	0/0/0	0.304230	0/0/0
10^{-4}	Interval	0.75	10	0/0/0	1.239804	0/0/0
10^{-4}	Support	0.25	4	0/0/0	0.015360	0/8/8
10^{-4}	Support	0.25	6	0/0/0	0.041729	0/0/8
10^{-4}	Support	0.25	10	0/0/0	0.138059	0/0/0
10^{-4}	Support	0.75	4	0/0/0	0.050317	0/0/2
10^{-4}	Support	0.75	6	0/0/0	0.221098	0/0/0
10^{-4}	Support	0.75	10	0/0/0	0.876580	0/0/0

Prediction inputs and all forecasts were saved before the evaluator opened the existing direct and spectral targets. Six synthetic tests checked polynomial identities, continuous suprema, the exact two-node radius, the factor-two transcript budget, first-log transport and the gradient premise. Numeric work took 7.46 host-CPU seconds on the A40 server with one BLAS thread; no training or GPU work was performed.

W PROSPECTIVE RECOVERY ON DISJOINT RIDGE COHORTS

Question and primary finding. Three early logs resolve all eight warm and all eight cold comparisons with matrix recovery at the primary tolerance, whereas scalar recovery abstains on every comparison. This extends the coefficient-family comparison to disjoint learning instances, but does not establish rejection value at that tolerance: every primary target is a true PASS. The restricted spectral interface also abstains throughout. Tables 24–26 report all prespecified tolerances and both paired states.

Fixed learning instances and observations. Exactly eight new roots use 16,384 training rows disjoint from the earlier 16,384 rows. A fixed outcome-blind permutation assigns eight children of 256 rows to each root. The CIFAR-100 source, frozen ResNet18 features, orthonormal 512-to-64 projection, global feature-only normalization, and 10,000 unlabeled probe rows are unchanged. All 100 one-hot output coordinates are retained. Each child minimizes $F_k(W) = \|X_kW - Y_k\|_F^2/(2 \cdot 256) + \gamma\|W\|_F^2/2$, with $\gamma = 1/1024$, mass $1/8$, and no bias coordinate. Warm starts use the fitted union ridge solution; cold starts use $W_0 = 0$. They are paired conditions on eight roots, not 16 independent repetitions. Full-batch, stateless gradient descent uses $\eta = 0.75/\max_k \lambda_{\max}(H_k)$ and target horizon $L = 6$. Only mass-averaged local displacements at steps 1, 2, and 3 are published, rounded per aggregate coordinate to spacing 10^{-8} with fixed rounding and numerical allowances. The target D_6 is the mean local endpoint minus the union endpoint. Its score is $s_6 = \|X_{\mathrm{probe}}D_6\|_F/\sqrt{10{,}000}$, with no additional division by $\sqrt{100}$. The primary $\tau = 0.01$ and secondary $\tau \in \{0.001, 0.05\}$ were fixed before target access; they are illustrative score-space tolerances.

Matched recovery and restricted information. The scalar and matrix families receive the same Hessians, masses, initial state, probe covariance, and $q \in \{0, 1, 3\}$ logs and allowances. Both use the same relaxation $B_k = C_k^{1/2}V_k$, $\|V_k\|_F \le 1$, and minimize the same block-normalized Frobenius residual plus ridge penalty: scalar blocks are $Q_m = a_mI$, while matrix blocks are unrestricted 64×64 matrices. The fitted-head residual row is retained for warm starts and omitted for cold starts, so cold $q = 0$ contains no observations. The unchanged grid comprises the zero operator and ridge ratios $0, 10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1$, with the fixed 10^{-12} zero-ratio singular-value cutoff. Intervals intersect candidates within a family, either at exactly q or through q, with identical error accounting. An upper endpoint at most τ gives PASS; a lower endpoint above τ gives FAIL; otherwise the decision is ABSTAIN. The restricted interface receives spectral support, norm bounds, and early-discrepancy budgets, but no directional logs or full matrices. It uses depths 1–3, a continuous interval as primary support, and finite numerical support as a fixed sensitivity analysis. Its one-sided bound permits PASS or ABSTAIN, never FAIL.

Table 24: All prespecified tolerances and paired states. Truth is PASS/FAIL/AMBIGUOUS; recovery entries are PASS/FAIL/ABSTAIN counts out of eight. The primary tolerance is $\tau = 0.01$. Direct and cumulative suites have identical decisions; their low-order interval endpoints need not coincide.

State	τ	Truth	Scalar $q = 0$	$q = 1$	$q = 3$	Matrix $q = 0$	$q = 1$	$q = 3$
Warm	0.001	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	7/0/1
Cold	0.001	7/1/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	7/1/0
Warm	0.01	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	8/0/0
Cold	0.01	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	8/0/0
Warm	0.05	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	8/0/0
Cold	0.05	8/0/0	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	8/0/0

Table 25: Every root–state comparison, with $q \leq 3$ intervals evaluated at $\tau = 0.001$. Scores and interval endpoints are in units of 10^{-3}. The last column is the matrix decision at that secondary tolerance; scalar recovery abstains in 16/16 rows. All scalar lower endpoints are zero. At the primary tolerance, 16/16 cases are true PASS and 16/16 are matrix PASS. Decisions use unrounded values and fixed numerical allowances. Root labels follow the frozen membership order.

Root	State	Actual score	Matrix lower	Matrix upper	Scalar upper	Decision
1	Warm	0.677918	0.644113	0.711552	147.719843	PASS
1	Cold	0.769510	0.729819	0.809190	147.456330	PASS
2	Warm	0.667723	0.634723	0.700675	138.089499	PASS
2	Cold	0.728684	0.690524	0.766958	138.238716	PASS
3	Warm	0.822427	0.785237	0.859760	150.480377	PASS
3	Cold	0.934434	0.890347	0.978470	150.632867	PASS
4	Warm	0.650976	0.601184	0.700676	137.572909	PASS
4	Cold	0.718629	0.659384	0.777741	137.499425	PASS
5	Warm	0.984655	0.945985	1.023339	150.292605	ABSTAIN
5	Cold	1.079725	1.034138	1.125171	150.450406	FAIL
6	Warm	0.671755	0.634768	0.708713	142.004760	PASS
6	Cold	0.758190	0.714529	0.801720	142.164187	PASS
7	Warm	0.769738	0.735131	0.804335	136.958726	PASS
7	Cold	0.859112	0.818410	0.899937	137.075583	PASS
8	Warm	0.653220	0.612091	0.694399	139.590728	PASS
8	Cold	0.729664	0.680591	0.778729	139.721155	PASS

What the outcomes establish. At the primary tolerance, the matrix-only advantage is eight correct decisions out of eight in each state; neither family resolves a comparison with zero or one log. The prespecified warm replication criterion (at least four net additional correct decisions) is met. The cold-rejection criterion (at least one matrix-only correct FAIL) and the requirement for both true primary regimes are not met. Restricted depth three adds zero correct warm PASS decisions over depth one, below the prespecified four-of-eight criterion. These three failed primary claims remain closed. At $\tau = 0.001$, matrix recovery gives seven PASS and one ABSTAIN in warm starts, and seven PASS and one correct FAIL in cold starts; scalar recovery still abstains throughout. At $\tau = 0.05$, matrix recovery gives eight PASS per state and scalar recovery again abstains. Direct and cumulative decisions coincide everywhere; their $q = 0$ and $q = 3$ interval endpoints also coincide, although some $q = 1$ endpoints differ.

The secondary threshold illustrates the decision boundary. Root 5 has warm score 0.000984655, with matrix interval $[0.000945985, 0.001023339]$ straddling 0.001, and cold score 0.001079725, with lower endpoint $0.001034138 > 0.001$. Thus its warm decision abstains and its cold decision correctly fails. This root is highlighted after observing the outcomes, not a separately preselected test; Table 25 includes every case. The secondary rejection does not rescue the failed primary rejection or both-regime claims.

Numerical validation, cost, and scope. Predictions for every case were frozen before computing any six-step target. The saved evaluations report coverage for all 864 matched candidates and 96 restricted depth/support checks, with no false definite decision among 576 suites; all 48 recorded

Table 26: Restricted-interface score-bound ranges across eight roots. Each depth uses the cumulative minimum through that depth. The interval is primary and finite support is the fixed sensitivity analysis. Both modes return 0 PASS and 8 ABSTAIN in every state, depth, and prespecified tolerance; FAIL is unavailable by construction.

State	Spectral information	Depth 1	Depth 2	Depth 3
Warm	Interval	[3.72081, 3.79134]	[0.95117, 0.96816]	[0.17653, 0.17930]
Warm	Finite support	[3.72081, 3.79134]	[0.29308, 0.33202]	[0.07379, 0.10387]
Cold	Interval	[3.34488, 3.41118]	[0.85593, 0.87167]	[0.16168, 0.16503]
Cold	Finite support	[3.34488, 3.41118]	[0.26410, 0.29968]	[0.07056, 0.09739]

prefixes satisfy their original allowances, and direct-gradient and spectral targets agree within the fixed allowance in all 16 cases. These are Float64 analytic-bound evaluations, not formally outward-rounded certificates. Numerical execution took 25.9754 seconds on one CPU of an A40-equipped host, with one BLAS thread, no GPU use, and 234.59375 MiB sampled peak resident memory. This is neither an encoder-training cost nor a speed comparison. The result concerns these fixed coefficient families and information interfaces on shared frozen features; it establishes neither scalar information impossibility nor globally optimal recovery, and makes no new accuracy, privacy, or production-logging claim.